Connectionist Modelling of Category Learning

Thesis

Connectionist Modelling of Category Learning

Paul D Bartos

Thesis submitted in partial fulfilment of the requirements for PhD

September 2001

Paul David Bartos BSc (Hons), MSc

Connectionist Modelling of Category Learning

Thesis submitted for the award of Doctor of Philosophy in the Psychology discipline of the Open University

18th September 2001

Abstract

A shortcoming is identified with respect to the ability of exemplar-based connectionist models of category learning to offer accounts of learning about stimuli with variable dimensionality. Models which may simulate these tasks, such as the configural-cue network (Gluck & Bower, 1988b), appear to be unable to accurately simulate certain data well simulated by exemplar-based models such as ALCOVE (Kruschke, 1992).

A task in which the advantage of ALCOVE is exemplified is the prediction of human learning rates on the six category structures tested by Shepard, Hovland, and Jenkins (1961). The ability of ALCOVE to simulate the observed order of difficulty depends on its incorporation of selective attention processes (Nosofsky, Gluck, Palmeri, McKinley, & Glauthier, 1994). This thesis focuses on developing configural-cue network models which incorporate these processes.

Informed by an information-theoretic approach to modelling the implementation of selective attention using a configural-cue representation, five connectionist models are developed. Each is capable of predicting the order of difficulty reported by Shepard *et al.* (1961). Two models employ a modular structure, but analysis suggests that these may lack much of the functionality of the basic configural-cue network. The remaining three incorporate dimensional attention schemes. These models appear to offer superior generalisability in relation to the simulation of learning about variable dimensionality stimuli.

This generalisability is examined by applying a variant of one of these dimensional attention models, to data collected by Kruschke (1996a) on the inverse base-rate effect and base-rate neglect. The model provides a qualitative fit to this data.

The success of these configural-cue models on these two tasks, only successfully modelled previously using two distinct types of representation, indicates that the approach has some potential for further applications. Differences between the models applied, however, indicates that more sophisticated conceptions of the attention process may be required to allow further generalisability.

Contents

Chapter 1: Introduction

1.1 Prolegomenon

The experimental investigation of category learning may be regarded as a continuation of research begun in the early history of psychology into 'associative learning'. The methods employed in the type of category learning experiments discussed in this thesis generally extend from the basic instrumental learning paradigm.

In instrumental learning experiments an animal (e.g. a dog) only gets reinforced (e.g. given food) if they produce the 'correct' behaviour (e.g. press lever) in the presence (or following presentation) of a particular stimulus (e.g. a bell ring). The dog may only get food *if* they press the lever *and* the bell has just rung. The expected pattern is that the appropriate stimulus (bell), becomes a discriminative stimulus, in that it gains control over the response (press lever). The animal may be said to have learned that if it hears the bell and presses the lever it will get food. It is discriminative in that if another stimulus is presented during the experiment (e.g. a flashing light) and is only followed by reinforcement if a different response is produced, (e.g. barking), then the light's presence enables discrimination between types of event which reward barking, and types of event which reward pressing the lever.

These two types are, broadly speaking, categories. Adding other stimuli to the training, with reinforcement given if and only if the 'correct' response (from barking or lever pressing) is produced, will increase the membership of these categories. While this may make them look more like 'proper' categories and less like stimulus-response chains, the distinction, at this point, is arbitrary. The category is defined as consisting of those stimuli associated with a particular rewarded response. It could also be described as the set of stimulus-response pathways which end at the response (e.g. barking), and begin at the stimuli positively associated with rewarded occurrences of the response.

To learn about learning, the basic paradigm may be extended in a variety of ways. One might, for example, alter aspects of the task to see if variations in the task produce reliable variations in the learning rate. This may be in terms of the rates at which errors are made given each variant of the task. This measure addresses the idea of task difficulty. The difference in difficulty between one task and another is a function of the differences

between the tasks and the nature of the system learning the task. A theory about the difference between the difficulty of two tasks is a theory about what is relevant to the cognitive system upon which task learning or performance depends.

Other experiments attempt to probe the nature of the knowledge which supports performance on a task. These experiments might involve the use of concurrent learning tasks. Here, following training on one task, the participants may be expected to learn a different task to gauge what effect existing knowledge has on performance of a new task. Alternatively, insights into the way in which the knowledge required for task competence is stored or represented may be gained by presenting experimental participants with new stimuli, in order to examine how they deploy or generalise existing knowledge to the new stimuli.

The categorisation experiments, which are the principal focus of this thesis, do not generally involve any specific methods of reinforcement. Feedback is usually in the form of some representation of the experimenter-defined correct response, for example, the category label. The basic models of associative learning developed to describe animal learning data, however, appear to have a high level of applicability to descriptions of human category learning. Consequently, similar models of learning are used in animal learning research and the study of human category learning. The connectionist approach to simulating category learning is the clearest example of this associationist heritage.

Connectionist models of category learning generally incorporate multiple theories about the communication processes and structures that might underlie learning and performance in categorisation tasks. These theories are implemented in terms of 1) rules that describe the way in which the network's connections alter, 2) models describing what it is about stimuli which get represented in the learning process, and 3) methods of determining the network's output.

These various theories are usually 'combined' to allow a model to simulate data from a particular experiment. Having been so developed, the model may be applied to related tasks, in order to evaluate the ability of the model to generalise its particular mode of description across tasks.

Attempts to generalise connectionist models to other tasks may inform developments of the model, or the theoretical assumptions or components underlying the

model. Obviously, the ability of a model, developed to account for one set of data, to simulate another set depends, to a great extent, on the similarity of the two tasks. The characteristics of connectionist models, like those of other kinds of model, depend largely on the experiments they are designed to represent.

One possible advantage of connectionist models over other modelling approaches is that, in some cases, the multiple theories they incorporate may be changed independently of one another. Problems with simulating data from a new experiment, for example, may be rectified by altering the way in which the stimuli involved in the experiment are represented by the model. The learning rule used may be the same in the two models.

This allows a certain amount of parsimony when it comes to model design and alteration, in that the functional components of the model may be systematically tested and evaluated in relation to one another. It can, however, lead to a high level of complexity in interpreting exactly which parts of the model are essential to its ability to simulate data. The components of these models are highly interactive in nature. The more components one introduces to a model, the less straightforward it is to identify what exactly it is about the model that is responsible for its successes and failures.

Alterations made to allow a model to simulate data from another experiment, however, should not necessarily prevent it from being able to simulate the data for which it was originally developed. If they do, the modifications should be accompanied by some theory as to why different models are required to simulate performance on the different tasks. Modifications to models should increase their generalisability, or should involve some principled account as to why the two experiments cannot be explained in terms of the same model. A failure to offer this generalisability, or a principled account as to why such a general model may be impossible, identifies a worthwhile challenge for research.

An example of the kind of challenge described above provides the main direction for the research described in this thesis. Shepard, Hovland, and Jenkins' (1961) investigation of category learning rates has turned out to be a particularly diagnostic test for models of category learning. It has also resulted in a situation in which the only models which are currently capable of simulating their findings are apparently incapable of being able to simulate fairly basic data from other categorisation and associative learning tasks. No *psychological* theory has been proposed as to why this is the case.

Shepard *et al.*'s (1961) study (partially replicated by Nosofsky, Gluck, Palmeri, McKinley, & Glauthier, 1994) involved examining the rates at which particular experimenter defined category structures were learned by human participants. The structures used the same stimuli; eight objects which could be differentiated from one another on the basis of variation in three dimensions. Each structure had two categories, containing four of the objects each. The tasks differed from one another in terms of the way in which the objects were divided into two groups.

The category structures used in the experiments will be described in detail in chapter 2. Shepard *et al.* (1961) however, noted systematic differences between the rates at which each type of structure was learnt. The order of difficulty for the structures could not be accounted for, at the time, by any model of associative learning, incorporating any model of stimulus representation. These authors proposed that it *might* be accounted for by a particular model of stimulus representation, now known as the exemplar model. This model, however, would require some secondary learning process, which they referred to as 'abstraction' or 'selective attention'.

They suggested such a process may allow the learning process to take advantage of the fact that some of the tasks had irrelevant dimensions. Task difficulty could be described as being inversely proportional to the number of dimensions which were irrelevant to perfect performance. Task difficulty described only in terms of associative relationships obtaining between stimulus representations and the category labels would not be sufficient.

A formal model of the way in which selective attention might allow an exemplar based model of category learning to account for Shepard *et al.*'s data was presented in 1984, in the form of the generalized (*sic.*) context model (or GCM) (Nosofsky, 1984, 1986). In 1992, a connectionist implementation of the GCM was developed capable of representing selective attention as dependent on a secondary learning process (Kruschke, 1992). This model provided quantitative fits to learning data produced in Nosofsky *et al*'s (1994) replication of Shepard *et al's* (1961) experiment which were superior to those produced by other models of category learning.

Despite the success of the exemplar approach on this particular task, the model of stimulus representation it employs is difficult to generalise to other tasks. One major

weakness is that it is unable to offer any principled, plausible account for experimental data relating to learning about stimuli that have different numbers of dimensions or components. This prevents it from being able to represent some of the most basic of associative learning phenomena.

This thesis, therefore, attempts to develop connectionist models under the constraint that they not only be capable of simulating Shepard *et al.*'s (1961) data, but that they are also able to simulate learning about stimuli with different numbers of dimensions or components.

1.2 Structure of the thesis

Chapter 2 provides a detailed description of Shepard *et al*'s (1961) category learning tasks and describes their findings as well as those of Nosofsky *et al*'s (1994). The diagnosticity of the data, with respect to attempts to produce models capable of accounting for them, is discussed and reasons for why some models 'fail' are described.

Chapter 3 provides a review of the ways in which major components of connectionist models of category learning have been developed. The first section describes the development and properties of basic learning and choice rules commonly used in these models. Their origins in associative learning theory and the early 'stochastic' models of learning is discussed. The shortcomings of these approaches, in terms of their lack of any specific representations of stimuli, are described.

The second section of chapter 3 provides an overview of various models of stimulus representation. The importance of generalisation in any model of learning is identified and the way in which this is effected by different models is discussed. The configural-cue form of representation is identified as emerging from basic associative learning theory as a means of accounting for basic learning phenomena. The origins of the exemplar approach in models used to account for psychophysical data is examined, and the differences between the assumptions required for its use in category learning and psychophysics is explored. The shortcomings of both the basic configural-cue model and the basic exemplar approach are described in relation to Shepard *et al.*'s (1961) findings and other observations.

The third section describes various approaches to the implementation of selective attention in connectionist models. Experimental findings suggesting that such a process

may be required are described. The implementation of selective attention in the exemplar network is described in detail with particular reference to the way in which it facilitates successful modelling of Shepard *et al.*'s findings. Attempts to implement selective attention processes using other forms of representation are described. Modular approaches, which involve structural organisation of stimulus representations into 'spatial' groups, are examined in relation to the data they can simulate, and the data they cannot.

A more recent class of model based on the idea of 'rapid attention shifts' (Kruschke, 1996a) is also discussed. Its applicability to the simulation of data concerning base-rate effects is described in detail as these experiments involve stimuli with different numbers of components. Significantly, these are experiments that the exemplar approach cannot simulate.

Chapter 4 deploys an information theoretic analysis of the Shepard *et al.* (1961) category structures. This approach is used to describe the relationships that obtain between a configural-cue model of representation and the category structures. The approach is used to develop three augmented configural-cue models. Each provides a way of describing learning in terms of the average rate of increase to the transmission rate of a proposed channel between spatially organised modules and a decision process. Two of the models use 'modular' weights to implement some form of selective attention, while a third uses a 'dimensional' model of the process.

The third model also employs a 'sequential sampling' process to represent the rate at which the various stimulus representations become active during a learning trial in terms of the operation of a Markov process. Selective attention, in this case, operates on dimensional sampling probabilities. All three of the models are capable of qualitatively simulating the order of task difficulty observed by Shepard *et al.* and Nosofsky *et al.* (1994).

In chapter 5 the simple modular models developed in chapter 4 are used to inform the design of two connectionist models. These models are also successful in offering a qualitative fit to the learning data of Nosofsky *et al.* (1994). The models appear to have numerous shortcomings with respect to their generalisation beyond the task modelled and these are described in detail.

Chapter 6 focuses on connectionist implementations of the simple dimensional attention model proposed in chapter 4. The sequential sampling model offers two distinct methods of implementing the dimensional attention process. One of these is based on modifying the sampling probabilities of the dimensions based on a back-propagation of error scheme. The second method involves altering weights between dimensions to control the sampling probabilities in terms of learnt relationships between the relative relevance of dimensions. Models making use of each method are described and tested. A third model extends the approach used in the second method to implement a rapid attention shift version of the process. All three models produce a qualitative fit to the Nosofsky *et al.* (1994) data. The shortcomings of the models are discussed in detail. These models, however, appear to offer superior potential for generalisation to other tasks than the models presented in chapter 5.

Chapter 7 applies a variant of the rapid attention shift model developed in chapter 6 to the modelling of two experiments investigating base rate effects. The model is capable of providing qualitative fits to the experimental data produced by Kruschke (1996a, experiments 1 and 3) showing base-rate neglect and the inverse base-rate effect. Significantly, exemplar models capable of simulating the difficulty of the Shepard *et al.* (1961) tasks cannot simulate these effects.

Chapter 8 discusses the extent to which the goals of the thesis, outlined in the first section of this chapter, are achieved by the research carried out. The further applicability of the models developed and tested in chapters 6 and 7 is also briefly discussed. While these models appear to offer ways of addressing the difficulties described in the previous section, the research indicates that the apparent flexibility of selective attention is not adequately captured. It is speculated that models using a single form of representation may be made to emulate the performance of other models, using different representations, by the operation of an appropriately specified selective attention process. Further research is clearly required into the flexibility with which stimulus information is used by learners to inform the development of more generalisable models of category learning.

Chapter 2: The Shepard, Hovland, and Jenkins (1961) Study on Category Learning Rates.

2.1 Category structure complexity and subjective difficulty

The 1961 Psychological Monologue, Learning and Memorization of Classifications (Shepard, Hovland, & Jenkins, 1961), was an account of an exploration of the effect of the logical structure of a classification task on the rate at which it could be learnt. The investigation was quite broad in scope attempting to address numerous issues including transfer of training; mode of stimulus presentation; differences between classification and identification; and rule formation.

Experiment 1, for example, looked at differences in learning rates for each of six experimenter defined task types (labelled I, II, III, IV, V, and VI). It also examined transfer of learning between successive repetitions of the same task type using different configurations of input, differences in learning rates between classification tasks and identification tasks, and the 'efficiency' of the rules developed by participants to carry out the classification tasks. The second and third experiments continued along these lines but explored different representations of the objects to be classified, memorisation of classification structures, and the efficiency of rules developed when presented with the entire structure at once with category labels attached.

The estimate regarding the relative difficulty of the six structures depends on the model one adopts regarding classification learning. Shepard *et al.* (1961) used two models to predict an order of difficulty. One of these was based on complexity of the rule required to determine membership without error. The other was based on a multivariate signal processing (McGill, 1954) analysis of the tasks. This analysis was interpreted according to an assumption that the more dimensions one had to pay attention to, before a reduction of the uncertainty regarding the category label to zero was possible, the more difficult the task would be (Shepard *et al.*, 1961, appendix). These models will be discussed in more detail below and in later chapters but they suggested an ordering of I, II, (III, IV, V), VI.

The results obtained, despite certain limitations, were highly influential in that they seemed to indicate a relationship between Shepard *et al.*'s measures of the complexity of the task and the difficulty of that task. Models of learning that prevailed in the field, at the

time, were based on stimulus generalisation and cue conditioning. These models either were incapable, in principle, of carrying out some of the tasks or, because they were not 'responsive' to the kind of complexity information, which seemed to predict difficulty, could not accurately simulate the observed difficulty.

This may be regarded as the basic finding and will be discussed in more detail below. The link between the measure of complexity and the difficulty of the task is represented throughout the results of the various experiments. The observation of the effect was repeated in a partial replication of the research, reported by Nosofsky *et al.* (1994).

The limitations of the original research principally involved the low number of participants (six in the first experiment, 20 in the second and third) which effectively renders the force of the results a product of their consistency across the various experiments. The results of individual experiments and what they say about things like transfer, different representations of the task structures, and rule formulation are, consequently, somewhat less significant.

Nosofsky *et al.* (1994) were motivated to overcome the limitations of the Shepard *et al.* procedures, with respect to the basic effect, in order to obtain data rich enough to enable quantitative evaluations of more recent models of category learning. Their results extended the Shepard *et al.* (1961) findings by collecting trial-by-trial data enabling average learning curves to be developed and the ease of learning for individual patterns to be identified.

2.2 Experimental structure for Nosofsky *et al.* (1994)

The experiment requires participants to learn the category assignments of eight, three-dimensional stimuli according to one of the six category types. Learning was trial and error for a fixed number of blocks or until criterion (whichever occurred first), with the participants being shown each object in random order, making their guess as to category label and then being given feedback in the form of the actual label.

Figure 2.1 illustrates the six basic category structures. The objects differed from one another in terms of independent variation in three dimensions. Each dimension consisted of a pair of substitutive features or attributes which, for illustrative purposes, are given as big-small, black-white, and triangle-square. The actual stimuli used by Shepard *et al.* (1961) consisted of images containing, at each of three positions, one of two

'thematically related' pictures (*ibid.* p.6). Each position represented a dimension such that one position might always contain a picture of a nut or a bolt, and another might contain a picture of a trumpet or a violin. Nosofsky *et al.*'s (1994) replication used geometrical shapes with lines that filled their interiors. These stimuli varied in terms of their shape (square or triangle), whether the internal lines were solid or dotted, and whether the shapes were large or small (*ibid.* p.354).

The different logical structures, shown in terms of a spatial representation in figure 2.1, are the six basic structures possible when dividing the eight objects into two mutually exclusive and exhaustive classes with four members each. While there are 70 possible arrangements of these objects ($8!/(4!)^2$) into two mutually exclusive sets of four there are only six basic types. Using the cube representations in figure 2.1, any of the remaining 64 possible structures can be obtained by rotations and reflections of one of the six cubes. These types will be described in more detail below.

A trial consisted of the participant being shown an object on a screen, being required to make a guess regarding its category label and push the appropriate button, and then being provided with feedback in the form of the correct category label. A block consisted of 16 trials in which each object was presented, in random order, twice. The first block was divided into two sub-blocks of 8 trials during which each object was presented, in random order, once. In the remaining blocks each object could appear at any two points of the sequence. Learning continued until participants had made no errors for four consecutive sub-blocks of eight trials, or had completed 25 blocks, whichever occurred first.

Each of the 120 participants learnt two category structures and were informed that the second problem was chosen independently of the first. Each pair of tasks occurred equally often, with the order balanced. Assignment of dimensions and physical values to the logical structure was randomised for each problem type.

As with Shepard *et al.* (1961) the idea was that differences in the rates of learning could only be attributable to differences in the category structure.

2.2.1 The category structures

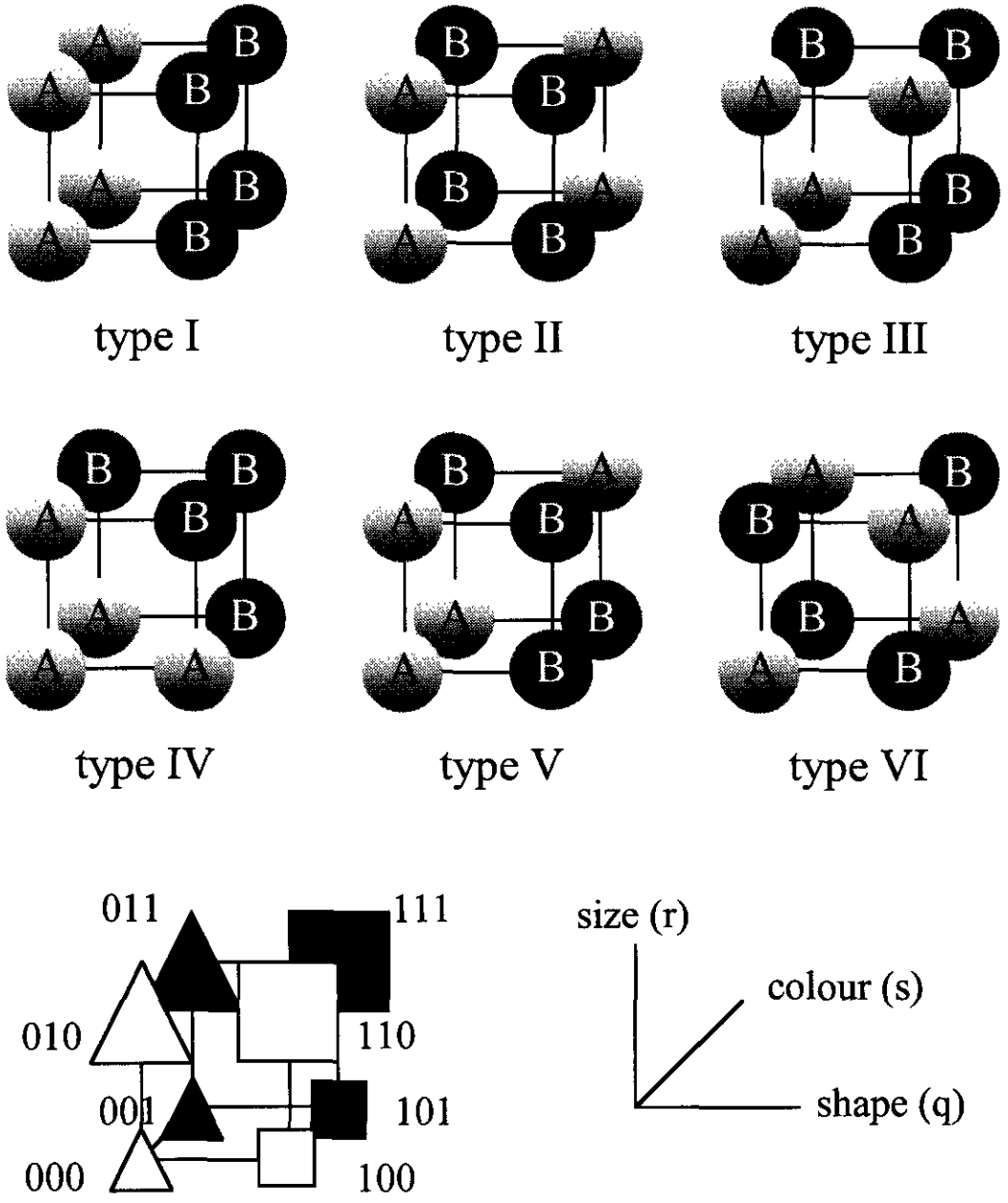

Figure 2.1: Abstract representation of the category structures used in Shepard *et al.* (1961), and Nosofsky *et al.* (1994). Category assignment given by letter A or B and white spheres or black spheres respectively. The binary co-ordinates of each object in the space (q,r,s) are shown next to each example object.

As can be seen from the cubes in figure 2.1, type I is a simple filtration task which requires knowledge of only one dimensional value (in the case shown, the shape of the object) to perfectly predict the category label. A rule for deciding category membership might state that if the object is a triangle it belongs in category A (and if it is a square then it belongs in category B).

Type II is an example of an XOR or condensation task. Knowledge of the values of two dimensions (shape and colour in figure 2.1) is required in order to predict the category label. The rule would state that if the object is white and triangular, or black and square, it belongs in category A (and if it is white and square or black and triangular then it belongs in category B).

Types III to V are somewhat different in that the requirement for knowledge of dimensional values may vary dependent on the object. They may be described as rule-plus-exception structures, for example in the type V structure all of the triangles except the large black triangle are members of category A and all of the squares except for the large black square are members of B. The differences between them are in the number of partially predictive dimensions involved in the structure. Type V has one such dimension (shape), type III has two (shape and colour) and type IV has all three.

There are different rules possible to describe membership but all require, to some extent, the involvement of all three dimensions. The rule-plus-exception structure described above is one. Another might involve, for the type V structure, the fact that all of the small triangles are in A and all of the small squares are in B, otherwise if it is a black square or a white triangle, it is in A, and if it is a white square or a black triangle, it belongs in B.

The latter rule is sequential in nature in that one can be certain of category membership by just knowing shape if, upon inspecting size one finds that the object is small. Half of the time only two dimensions are required. If the object is large then one is dealing with the top-half of the structure for type V which, being an XOR problem, requires knowledge of both dimensions, neither of which is size, hence all three dimensional values will have to be known half of the time. Adopting a rule-plus-exception representation requires knowledge of all three dimensions for every object presented in

order to determine whether or not it is an exception. The rule may be described fairly concisely, however, by stating the rule and enumerating the exceptions. These structures are discussed more below. Shepard *et al.* (1961) suggested that these tasks should be harder than the type II because they cannot be completely described in terms of the values of two dimensions.

They should, however, be easier than the type VI structure. All three dimensional values are required before any object may be classified correctly. A rule for this structure may require enumeration of each category member.

Alternatives are possible but do not seem to occur for experimental participants very often. One example from Shepard *et al.* (1961) involved sequential information. One participant stated the rule in terms of how many values had changed between presentations. If one value, or all three values change, then the object is a member of the other category. This rule emerged only in the transfer experiment where participants had to carry out repeated tasks of the same type. Alternatively one may recode the dimensional values to produce a parity problem (Nosofsky *et al.*, 1994, note 1). Parity problems will be discussed in the next chapter.

Nosofsky *et al.* (1994) also investigated the different learning rates at which patterns with different logical 'status' within the structures were learnt. Objects in types I, II, and VI are uninteresting with respect to this in that all objects have an equal logical status. This means that no particular object would appear to be any more difficult to categorise than any other. They expected that differences would emerge for different patterns in the type III to V category structures. Some of the stimuli seem to be more central to the category than others. Nosofsky *et al.* (1994) describe the differences between the stimuli in terms of a spatial metaphor and whether they are central, peripheral, or exception members.

A central member, according to Nosofsky *et al.* (1994), is one that 'always participates in the single-dimensional rule and is never considered an exception, whereas peripheral members will sometimes serve as exceptions, depending on which dimension is used for the rule.' (p. 356). Figure 2.2 illustrates the different roles of each object in category structures III to V.

It was expected that central members would be learnt with fewer errors than peripheral ones, and that the most errors should be made on exceptional objects (*ibid.*). The reason for this is dependent on the model used to describe the learning and representation process. Theories which suggest that responding to a stimulus is affected by the similarity of that stimulus to other stimuli, would predict that difficulty is a function of how similar an object is to members of its own category and its similarity to members of the other category. Central objects are obviously 'closer' to the other members of their category than peripheral ones, and so their category assignments will be easier to learn.

This 'stimulus generalisation' approach and some of its alternative manifestations will be discussed below and in more detail in chapter 3. Another interpretation of the difficulty is provided by a rule-based approach. Here, a 'central' member would be characterised by its ability to participate in more perfectly predictive, two-dimensional 'rules' than the other objects. Figure 2.2 illustrates these relations using arrows to indicate the direction of the other member of the 'redundant' pair.

Category A in Type III has two central members; object 000 may be a part of the perfectly predictive 'small triangles' or 'white triangles' rule; object 010 may be part of the 'white triangles' or 'large white' configural rules. The peripheral members only participate in one of these rules each. In type IV the difference is that there is only one central member per category. These participate in three perfectly predictive configural rules. The peripheral members, again, participate in just one each.

For type V the central object of category A falls within two perfectly predictive configurations, 'white triangles' and 'small triangles'. The peripheral objects are only located within one configural rule each and the exception is not located within any.

The difference in difficulty may arise from the fact that central members may be located in a rule more easily than peripheral or exception members. For type IV the central member is located in a rule which just requires the value of any two of the three dimensions. For types III and V, the situation is slightly more complex. The central members are located within rules that require knowledge of two dimensional values but for each central member, only two of the three possible configurations are relevant. Peripheral members require two particular dimensional values to be known for uncertainty to be

reduced to zero. Different peripheral members require different combinations of dimensional values.

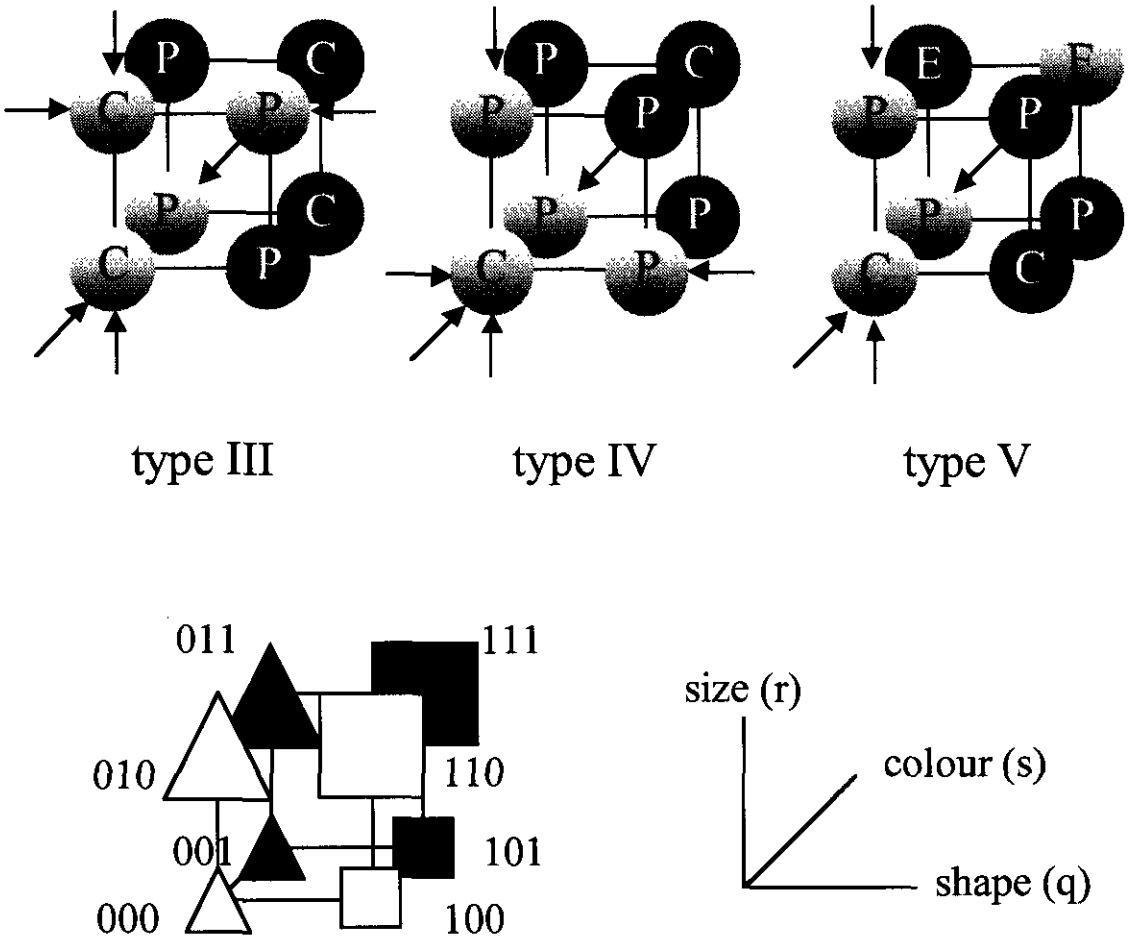

Figure 2.2: Category structures III to V with objects labelled according to Nosofsky *et al.* (1994) classification as central (C), peripheral (P), or exception (E). Arrows indicate the direction of the other object in a perfectly predictive configural rule within which the object may be located. Category A represented by white spheres, category B by black. Note, only arrows for category A are shown. See text for further analysis.

2.2.2 Results

The Shepard *et al.* (1961) study produced a wide array of results consistently showing the predicted order of difficulty for the six tasks. The Nosofsky *et al.* (1994) partial replication showed the same order of difficulty at high levels of significance for average number of errors and number of trials to criterion.

Like Shepard *et al.* (1961) the replication also noted an overall practice effect between the two tasks each participant was required to carry out. The ordering was the same for the two problems with the same significance. No significant difference was noted between types III to V. The data represented in figure 2.3 is based on error averaged across the two problems in the Nosofsky *et al.* (1994) replication.

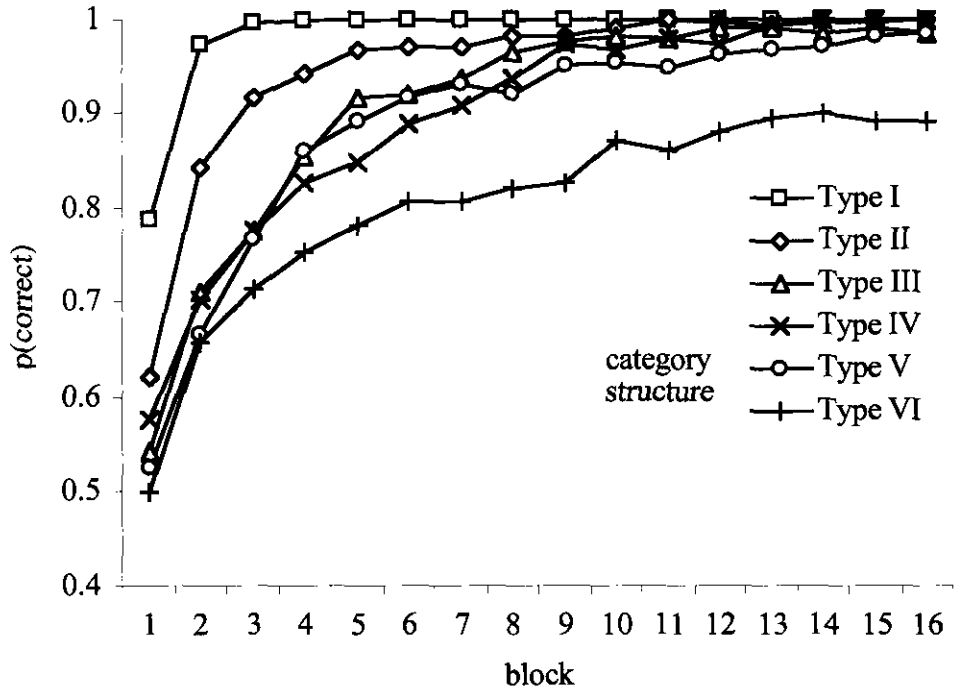

Figure 2.3: Mean probability of correct response per block (16 trials) of learning on each of the six category types from the Nosofsky *et al.* (1994) replication.

With regards to the individual pattern difficulties, Nosofsky *et al.* (1994) found the predicted order of pattern difficulty in tasks III to V. Central members were learnt with fewer errors then peripheral ones which were, in turn, easier to learn than exceptions.

As discussed briefly above, Shepard *et al.* (1961) explored a variety of tasks involving the same basic structures to examine differences in performance. The transfer of training between consecutive experiments with the same structure (or rather reflections and rotations of the same structure) is one area where data was produced. While interesting in seeming to suggest higher order learning processes, the relevance of the data is compromised somewhat by the limitations described above.

Much of the analysis made by the authors (*ibid.*) was focussed on the basic effect of the difficulty of the task, as a function of its complexity. This relationship was most clearly indicated in the Shepard *et al.* study in the context of the initial difficulty of the task types, i.e. the average difficulty of each type of structure when being learnt for the first time.

This aspect of the experiment may be the most salient with regards to offering clear constraints on modelling, as it marginalises the effects of different types of information on the task difficulty (e.g. familiarity). This was the focus of the Nosofsky *et al.* (1994) replication and analysis but, while relating to more sophisticated models, the issues raised by Shepard *et al.* were still the relevant issues, regarding the modelling of the data.

2.3 Discussion: Shepard *et al.*'s (1961) analysis of the first problem difficulties

The original importance of the results reported by Shepard *et al.* (1961) was two-fold. The first area relates to the fact that participants could, generally, reach criterion on all of the tasks. The second concerns the order of difficulty of each task.

2.3.1 Object representation and cue conditioning

As discussed in chapter 1, the study of category learning has its origins in research on instrumental learning. At the time of Shepard *et al.*'s (1961) study, learning paradigms were generally simple in nature, typically characterised as versions of a type I category structure. The models developed to describe learning rates for these simple experiments were, generally, as complicated as required by the particular data being modelled. This would involve the model only being able to represent the learning of associations between individual features, or cues, and the response. The learning rate for each feature-to-label association would be a function of the conditional probability of the label being present given the presence of the feature.

Shepard *et al.* (1961) suggested that this type of model may be realisable as a 'working' model in terms of a perceptron (Rosenblatt, 1958). However, they also pre-empted Minsky and Papert's criticism of this type of model (Minsky & Papert, 1969), by pointing out that these models would not be able to predict criterion, or even greater-than-chance performance on certain of the problem types in their experiment (Shepard *et al.*,

1961, p.30-32). This is discussed particularly with reference to the type VI structure, but is also true of any structure where no single cue perfectly predicts the category label.

The problems with these models relate to the way in which the object is represented. The learning model describes the way in which responses become connected with the *representation* of the object. The assumption had been that it was appropriate to represent the object in terms of the separate features of which it consists. Success in tasks where no individual cue predicts the category label, but combinations of cues do, suggest that this form of representation is not, by itself, sufficient to account for a wider range of human learning data.

Shepard *et al.* (1961) suggested that models based on the cue conditioning approach but incorporating unique representations of the whole stimulus and, possibly, representations of configurations of cues would in principle be capable of performing each task. At the time, however, the lack of formal rules to account for how these models might learn meant that it was difficult to predict their asymptotic performance. As will be described in chapter 3, this type of model, now generally known as the configural-cue network, was tested by Gluck and Bower (1988b) and it failed to simulate the observed order of difficulty.

2.3.2 Stimulus generalisation, selective attention, and complexity

Another representational solution suggested by Shepard *et al.* (1961) is to use the stimulus generalisation model, described briefly above. This model was developed to account for data from identification experiments (Shepard, 1957). These experiments involve the participant having to learn to produce a unique response to each of a set of stimuli. These stimuli vary in terms of the number of features they have in common with one another.

In order to account for identification data, the model suggested that the probability of confusing one stimuli for another, and consequently producing its response rather than the appropriate one, is a function of the number of features those stimuli have in common. The representations used are of the unique combinations of cues that describe each stimulus. Unlike the cue conditioning models, where representations are either present or absent for each object, the representations are 'present' or associable to varying extents, dependent on their similarity to the actual stimulus presented.

Like Nosofsky *et al.* (1994), Shepard *et al.* (1961) focussed much of their attention on the consistent ordering of difficulty noted for problem types when learnt for the first time. They specifically attempted to apply the stimulus generalisation model to the data.

The stimulus generalisation model predicts that the difficulty of a task could be indexed by the ratio of the average similarity of an input to members of the wrong category over the average similarity to members of its own category. Categories made of members with little in common would be harder to learn than categories with members with a lot in common.

The way in which similarity is calculated in these models will be described in detail in the next chapter, however, when applied to the category structures the order predicted by the model was not the same as that observed in the data. The prediction made of I<(III, IV, V)<II<VI (with differences fairly small between all types) is understandable given closer inspection of the category structures in figure 2.1. Categories in types III to V have members which are clustered together around their central members (see figure 2.2), whereas for type II the membership of each category is split into two clusters located on opposite corners of the cube.

Shepard *et al.* (1961, p.29) proposed that the reason for the failure of the model to predict the difficulties, was that it was unable to abstract, or selectively attend to, only those dimensions relevant to the task. The theory abstractly specified was attenuation of the extent to which differences in irrelevant dimensions mediate similarity judgements relative to differences in relevant dimensions.

This, it was suggested, may result in the appropriate level of error for each type being predicted by the generalisation model. This was later demonstrated to be the case by Nosofsky (1984) who showed that by using appropriate 'attentional' parameters for each dimension, for each task, an accurate prediction regarding the level of difficulty could be made using these stimulus generalisation models. The details of this will be discussed in the next chapter but the relevant issue that emerges is that the relevance of a dimension is *task dependent*. As such it is something that a model based on stimulus generalisation will have to acquire from experience with the task, i.e. learn.

Shepard *et al.* (1961) concluded that the models of learning available were either not capable of predicting the task difficulties or not sufficiently specified to make detailed predictions.

In an effort to quantify, in some way, just what it was about the problem structure that appeared to control difficulty, an interesting but somewhat different approach was taken. This approach took, as its starting point, the assumption that one could describe the tasks in terms of a binary search structure where the number of dimensional values required (assuming an optimal start of the sequence with no dimensions repeated), before the category membership may be unequivocally decided, corresponds to the length of that search (*ibid.* p.33).

This is one measure of complexity, in terms of the proportion of the total number of 'bits' available (three independent dimensions, with two equiprobable values each yields a total of 3 bits), required to make the decision, on average. This predicts that type I requires one bit, type II two bits, and type VI three bits. Because the types III to V need all three bits, but not all of the time and can sometimes be certain with just two, it can be said that these will have bit-rates of between two and three. In the appendix of the paper the authors attempted to define a measure of difficulty based on this 'length of search' model.

This was expressed in terms of the distribution of uncertainty-reducing information across the task's dimensions. Assuming the optimal start point, and shortest sequence length, one could determine how much uncertainty (from a total of one binary digit or bit) is removed about the category label given knowledge of one, two, and three dimensional values (Shepard *et al.* 1961, appendix).

The authors presumed that the difficulty of a task would depend on the number of dimensions one had to be simultaneously aware of to use the rule or extract the necessary information (*ibid.*). They suggested that a task would be more difficult as the amount of information in the second and third dimensions increased. They suggested an index of difficulty could be defined by weighting the information in each term by a coefficient proportional to the number of dimensions involved in that term and then summing the products.

With appropriate dimensional parameters the authors produced the correct ordering, but what this model says about the learning process is somewhat abstract. The

approach is returned to in chapter 4 in much more detail where an explicit signal processing approach to theories of learning and categorisation is described.

Since the original study, models of human learning and classification have developed significantly. The purpose of Nosofsky *et al.*'s (1994) replication was to gather data that was detailed enough to enable the quantitative evaluation of some of the more influential models. This concerned the ability of the various models to quantitatively fit the block-by-block average error rates for the six category structures. These models are 'full simulations' in that they implement specific theories about representation *and* learning, to show trial-by-trial changes in performance.

The relationship of these simulations to the problems involved in the Shepard *et al.* (1961) category structures is the subject of the next chapter. However, the assertions made that models may have to incorporate some means by which the complexity of the task mediates its difficulty, remain valid. The more modern analysis deals only with models that are, in principle, capable of performing the tasks. It therefore focuses on the extent to which task complexity is predicted by each model to affect difficulty.

Chapter 3: Learning and representation in models of categorisation

This chapter deals with the various theoretical components of category learning models. As will be discussed, there are two central components to any theory of learning, one is a theory regarding the way in which events become associated with one another, the other is a theory regarding what it may be about a particular event which gets associated with a particular response.

The development of models of learning illustrates this interaction of theories well. Models of learning have, as research has progressed, had to adapt in order to allow the predictions they make regarding certain stimulus sets to generalise to different types of stimuli. In many cases this has just involved the development of different means of representing the stimuli and applying the new representations to whatever theory of associative learning which was contemporary.

As theory has developed, however, the representations that appear to be the most appropriate for particular stimuli have suggested more complex learning rules. These rules, which may be described as incorporating ideas of selective attention to the process modelled, involve task dependent 'extraction' of relevant information from a given representation.

In this case the complexity of all aspects of the model increases as the representations must be 'suitable' to have relevant information extracted from them. The learning model must be able to generalise across cases where this process operates as well as cases where it does not. Learning, representation, and selective attention are described in turn in this chapter with particular reference to their relationships and, in some cases, inconsistencies.

The following is not an exhaustive account of models of categorisation and learning. It is generally restricted to those models which either are connectionist in nature or which can be, and have been, used to inform the design of connectionist models.

3.1 Learning theory and categorisation

An analysis of the development of learning theory is informative with respect to the way in which models of categorisation, which frequently involve learning, are constructed.

Early experimental work with animals revealed, as described in the previous chapter, that one could predict the responses of an animal to a particular stimulus using, most simply, information about the frequency with which the response to that stimulus was accompanied by reinforcement of some description.

This observation applies equally well to the learning of the more complex stimulus-response relationships examined in categorisation research. As will be described, however, as the experimental paradigms used to investigate theories of learning have developed, factors influencing this simple frequency approach have had to be adapted.

3.1.1 Early mathematical learning theories: Clark Hull and 'habit strength'.

Early attempts to formalise the relationship between a stimulus and a response generally consisted of attempts to derive '*the* learning function' (Bush, 1960, p. 126). This would involve developing a mathematical expression which captured the relationship between, say, the number of times a response to a stimulus had been reinforced and some index of the likelihood of the stimulus resulting in the same response if it were to be presented again.

An early pioneer in the area of mathematical models of learning was the researcher Clark Hull. His 1943 book, *Principles of Behavior*, consisted of an attempt to advance research in behavioural science by offering testable formalisations of the 'habit'.

Hull described a habit as an 'intervening variable', hypothetical entity, or logical construct. He noted that while there were obvious risks associated with the use of such constructs in any scientific endeavour, the potential payoff, as shown by other sciences such as physics, was enormous (Hull, 1943, p.22).

His concern was to develop mathematical models which could capture some of the basic data produced by experiments on animal learning or conditioning. The *habit*, for Hull was a 'persisting state of the organism' resulting from reinforcement (*ibid.* p. 102) which could be described in terms of a set of receptor-effector connections. He was not specifically referring to a particular mode of action or specific behavioural response as the term habit was broadened to include all of the possible reactions that might result from reinforcement. This would include not just an overt conditioned behavioural response but any other responses that might occur, for example galvanic skin responses.

Hull suggested that the habit could be described in terms of the notion of 'habit strength' with notation ${}_sH_R$ to indicate the strength of a habit, H, which consists of the production of response R when presented with stimulus S. He proposed that habit strength would increase with the number of reinforced responses in a way captured by the following function;

$$ {}_sH_R = M - Me^{-iN} \qquad (3.1), $$

where M and i are positive constants and N is the number of reinforced responses to the stimulus (*ibid.* p. 119). As figure 3.1 illustrates, i controls the *rate* at which reinforcement affects habit strength. The constant M controls the maximum habit strength.

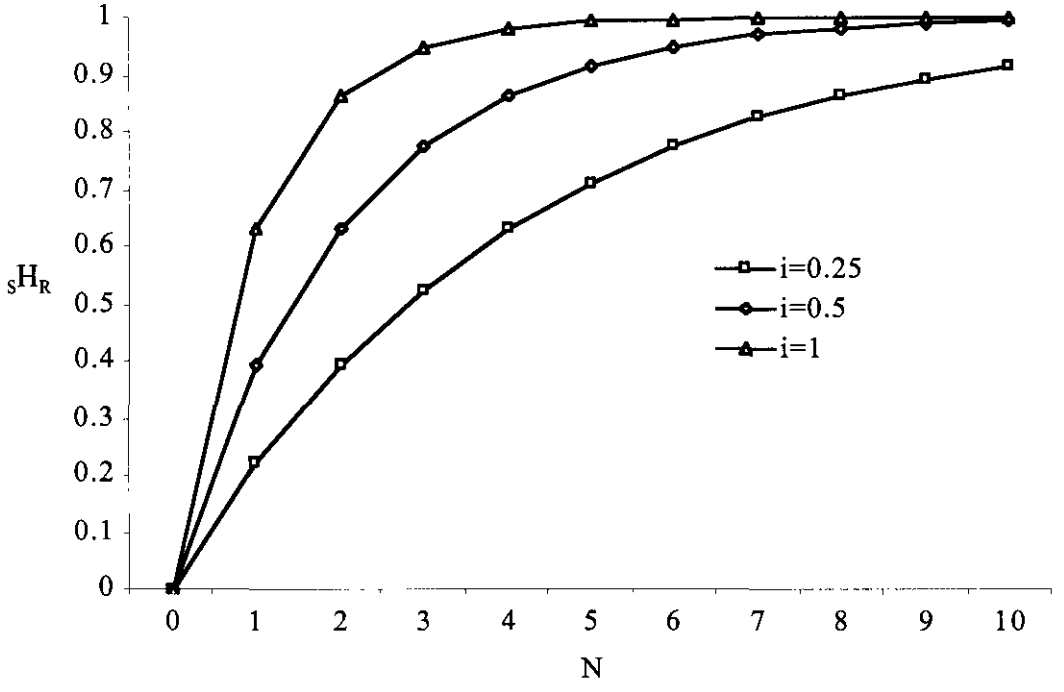

Figure 3.1: Graph of equation 3.1 showing habit strength, ${}_sH_R$, with increasing, number of reinforced responses, N, for three different values of constant i. The constant M is set to 1 for all graphs.

Hull, wanting to define a 'centigrade' scale of units of strength which he referred to as 'habs', set M to 100. The basic intuition behind the selected function was that the strength of a habit would be increased by some constant value, i, each time it was reinforced towards some physiologically determined maximum. As it approached this

maximum, the effect of reinforcement would diminish according to how close to the maximum the strength actually was.

This was based on observations concerning numerous measures of different types of response including 'autonomic' responses such as salivation and galvanic skin responses (where the maximum may be most clearly described as a physiological constraint), but also on behavioural action probabilities. His choice of the exponential function was actually a method of describing the curve produced by the 'positive growth function'.

The positive growth function was acknowledged by Hull to be one of many algebraic functions capable of producing a curve which could fit observed results. It was selected due to its usage at the time in fitting a number of empirical observations of various biological growth and decay processes (*ibid.* p.114). The function described in detail by Hull in the text, but not presented as a formal learning function, is as follows;

$$\Delta_S H_R = F\left(M - {}_S H_R\right) \qquad (3.2),$$

where F is referred to as the growth factor for a given reinforcement context and is a positive constant (*ibid.*). F is related to i in equation 3.1 but, rather than consisting of a constant increment, F determines the fraction of the 'unrealized potentiality' transferred to the habit strength (*ibid.*).

Hull, for some reason, regarded the formulation of equation 3.2 as 'rather clumsy' (*ibid.* p.119) and decided to use the exponential formulation given in equation 1 to describe the accumulation of habit strength as a function of number of reinforced responses. Equation 3.1 enables one to acquire a value of a particular habit strength at a given stage in the learning experiment, its form being particularly useful for evaluating the interaction of habits acquired across a known number of trials. The two produce identical curves when the value of i is calculated, as stated by Hull (*ibid.*) according to the following formula,

$$i = \log_e \frac{1}{1-F} \qquad (3.3).$$

Hull went on to address the factors which might influence the parameter M in the above functions. He identified and proposed formalisms of varying specificity to address such issues as the amount of reinforcement given, time delay between response and reinforcement, and the delay between the stimulus presentation and the response. These

factors are not particularly relevant to this present work and so will not be described in further detail.

Later chapters in *Principles of Behavior* looked at issues regarding compound stimuli, stimulus generalisation and probabilistic relationships between habit strength and behavioural responding. Unfortunately these issues, which came to assume greater significance in psychological learning research, seem to be somewhat disconnected from the learning functions described in Hull's work. It is thus not clear how all of the various components of his theories may be related to one another.

3.1.2 Stochastic learning models

Following Hull's pioneering work, mathematical modelling in the behavioural sciences enjoyed a rapid rise in popularity. The proliferation of models was influenced strongly by developments in the mathematical analysis of experimental data using statistical techniques such as analysis of variance and information theory. Owing to this contribution, mathematical approaches to learning became much more focused on developing models capable of specifically representing changes in the probabilities of particular responses being produced under controlled circumstances.

The successors to Hull's models were, in many ways, fairly similar although there was a marked change in emphasis. The psychology of the 1950's and 1960's had moved away from using the physiological and motivational 'constructions' within which researchers like Hull located their theories. These were, ostensibly, abandoned altogether in favour of simply constructing mathematical models capable of representing and predicting experimental data. As an alternative to the use of physiological terminology, however, new constructions were adopted related to the engineering mathematics and statistics deployed in the discipline.

One of the areas that had been particularly problematic was defining what was meant by 'reinforcement'. William Estes, commenting on the development of statistical learning theory, suggested that a shift from defining what reinforcement was, towards making quanititative statements about how it operates was justified (Estes,1959, p.404).

Sentiments such as this led to a generalisation of learning theory to more diverse experimental scenarios. The learning that occurred, for example, when a human participant was simply provided with one of two tones, dependent on their making an experimenter-

defined correct or incorrect response, was describable using similar basic 'rules' to the rules which might be used to describe the learning taking place in a rat being shocked for taking the 'wrong' turn in a maze. The characteristics of the learning were, therefore, separated from the reasons *why* this learning was taking place at all.

The basic observation of Hull and other early researchers, that the probability of making a particular response increased as a function of the number of occasions on which this response was reinforced, remained a basic tenet of learning theories. The emphasis, however, shifted from the representation of specific stimulus-response relationships to one of describing the changes in response probabilities as a function of the statistical properties of the stimuli used in the experiment.

The model of the decision and response aspect of the process was one of a stochastic system. The response produced was only probabilistically related to differences in whatever mediating pattern of response strengths one might care to propose. Learning could be described in terms of the changes in response probabilities which occurred when a response was reinforced or not.

The models produced by researchers such as Bush and Mosteller (e.g. Bush & Mosteller, 1955, Bush, 1960), Estes (e.g. Estes, 1959, Atkinson & Estes, 1963), and Restle (Restle, 1955), could be similarly applied to predicting asymptotic performance after a certain number of trials or for computing trial-by-trial increments in response probabilities.

Despite differences between the above researchers regarding the assumptions underlying the learning functions, the basic formulations were pretty much the same as each other, and also fairly similar to Hull. Bush (1960, p.132) gives the following 'rule' to describe the change in probability of a response class with probability p, where there are two response classes with probability assumed to be p and 1-p,

$$p_{n+1} = \alpha_i p_n + (1-\alpha_i)\lambda_i \qquad (3.4).$$

Here n is the trial number, $(1-\alpha_i)$ is a rate parameter and λ_i is the asymptote of learning in a given experiment for this response. The function may also be expressed in terms of the change in response probability as follows,

$$\Delta p = (1-\alpha_i)(\lambda_i - p) \qquad (3.5).$$

In the case of these models the subscript i refers to the response and as such there is no obvious representation for the stimulus in the equation. The same is true of Estes' stimulus sampling model. Estes (e.g. Estes, 1959) represented the function as,

$$\Delta p_i = \theta(1 - p_i) \qquad (3.6).$$

In this function the asymptote is assumed to be 1 for the response i and the $(1-\alpha_i)$ is replaced by the parameter θ which is referred to as the stimulus-sampling rate.

In effect $(1-\alpha_i)$ and θ are doing all the work of representing the stimuli involved in the experiment. For the Bush and Mosteller model α_i was a measure of the ineffectiveness of the stimulating event at altering response probabilities (Sternberg, 1963, p.22). Thus the higher its value, the lower the learning rate. It was conceived as, basically, a function of some measure of the average similarity between the stimuli (generally an array of stimulus elements) presented when one response was correct to the stimuli presented when the other was correct. The stimulus sampling rate, θ, relates to the proportion of relevant cues sampled on a trial. The assumption is that stimuli become conditioned in an all-or-nothing way to the reinforced response on each trial, dependent on whether they are perceptually sampled during that trial (Estes, 1959, p. 399). The θ refers to the probability that relevant stimuli are sampled, or used to inform the decision, on a given trial.

The rate parameter used by Restle (1955) was also called θ and was, he suggested, related to the proportion of relevant and irrelevant stimuli, or 'cues' as he referred to them. Where r is the number of relevant cues on a trial and i is the number of irrelevant cues on that trial Restle proposed a simple relation,

$$\theta = \frac{r}{r+i} \qquad (3.7).$$

This measure is, again, similar in intent to the stimulus sampling rate and the $(1-\alpha_i)$ figure used in the above equation. All of these measures are intended to represent, in some way, the similarity of the stimulus arrays, with the intention of operationalising the idea that learning is faster when the stimulus presentations requiring different responses are easy to differentiate from one another.

Unfortunately one problem remains in that θ is a parameter that actually must be estimated from experimental data, specifically using error in performance figures. Restle's use of equation 3.7 was, as pointed out by Bush (Bush, 1965, p. 171), somewhat

unjustified as, in reality this parameter had to be estimated. The power of these means of characterising the stimulus set, lies in the fact that once theta or alpha was estimated from error data for one experiment, its value could be estimated with some success as a function of the relationships between stimuli on similar experiments.

The above models are notable for the fact that they seem to characterise the stimulus set as a whole and attempt to describe its role in terms of single variables. They differ from Hull's treatment of the process as the development of a stimulus response connection in that they do not tend to address individual stimuli at all.

The problems of estimating the parameters involved in terms of relationships hypothesised between different experiments according to the relationship between stimulus sets are both practically and theoretically considerable. With hindsight, these issues may have been more readily addressed by paying more attention to the individual relationships between stimuli and responses.

The problems of the above models would be further exacerbated if one was attempting to apply them to *concurrent* learning tasks in which the characteristics of the stimulus set are changed at some point during training. Here the results of such a change are the data of interest. In this case some method of representing the relationship between individual stimuli and the response probabilities would be required to model any effects.

3.1.2.1 Luce's identification choice model

An early model which, in a way, addressed this issue was given by Luce (1963) based on a model published in 1964 by Bush, Luce, and Rose (1964), in the context of a psychophysical learning model for complete identification tasks. Assuming that j is the correct (determined by the experimenter) response to stimulus i and the conditional probability of j on trial n given i is $p_n(j|i)$ then the probability on trial n+1 given presentation of some stimulus k is,

$$p_{n+1}(j|i) = p_n(j|i) + \eta(i,k)\theta_k\left[\delta_{jk} - p_n(j|i)\right] \qquad (3.8).$$

Here $\eta(i,k)$ is a measure of the similarity of stimulus i to stimulus k. It equals 1 if i=k. The parameter θ_k is a rate parameter applying when k is present and δ_{jk} is known as the Kronecker delta which equals 1 when j=k and zero otherwise, (Bush *et al.* 1964, p. 211). The Kronecker delta, in this context, means that the conditional probability increases if j is the correct behavioural response, its value being 1 in this case. It decreases if j is not

the appropriate response given input k. The conditional probabilities of j given i, summed across all j thus remains at one.

The model is a means of describing the nature of a distribution of response probabilities across a space or continuum in, or along, which stimuli may be located relative to one another. The training 'shapes' this distribution. Stimuli are assumed to be related to one another in terms of some measure of the distance between them on some experimenter-determined scale. No specific representations of the stimuli are present in this model.

The model predicts that the conditional probabilities of different responses given different stimuli alter as a function of the similarity of the stimulus to the one that is actually being presented. An important property of the model is that it enables one to calculate the asymptotic conditional probability of a response given any stimulus in terms of its similarity to *only* the stimuli that have been presented. This measure was also developed by Shepard (1957) as a means of predicting confusion errors. For the identification task the expected conditional probability of producing response j when presented with stimulus i as the number of trials, n, approaches infinity is as follows (from Bush *et al.* 1964, p. 211),

$$\lim_{n\to\infty} E\left[p_n(j\,|\,i)\right] = \frac{\eta(i,j)b(j)}{\sum_{k=1}^{m}\eta(i,k)b(k)} \qquad (3.9).$$

The parameter b(k) is a response bias where b(k) =p(k) θ(k), where p(k) is the probability of the response across the experiment and θ(k) is the learning rate for that response as shown in equation 3.8 (*ibid.*). Taking probabilities of different responses and their learning rates to be equal, the equation reduces to the similarity of the input i to the stimuli j, for which j is the correct response. This is divided by the sum of the similarities of the stimulus i to all stimuli k for which k is the correct response (including j).

This has contributed to the idea that the capacity to associate stimuli with responses emerges from 'detectors' of some description for items that *have* actually been presented. The model will be described in more detail in the section on stimulus representation, below, as it has proved to be highly influential in the field of category learning.

The increment function is, however, a significant development on the Hull model. By including the parameter $\eta(i,k)$ control is enabled over the level at which association develops between a particular stimulus and a response according to the extent to which the detector or 'representation' is 'activated' by the current stimulus.

While Hull acknowledged that stimulus energy would probably have some role in the accumulation of habit strength (Hull, 1943, p.181) he failed to include it in his model describing the increment. As such, the inference may be that all connections or habits contributing to the effective reaction potential, even if made active only as a result of stimulus generalisation, would be conditionable at an equal rate.

3.1.2.2 Response probabilities and choice functions

Another advance on the Hullian approach is the inclusion of a simple function to determine choice probabilities. While Hull incorporated ideas of stimulus generalisation into his work, his method of handling 'conflict' between competing responses was somewhat awkward to interpret.

The Luce (1963) and Bush *et al.* (1964) model is specifically a model for generating *choice* probabilities. This contrasts, to some extent, with earlier stochastic models which generated *response* probabilities. In these models the rules described above applied to situations where there were two responses, with p(i) being calculated and the other assumed to be 1-p(i). For larger response sets one might have to suggest that each had its own probability with the outcome, or choice, being determined by normalised probabilities.

The determination of choice probabilities, in terms of the sum of evidence in favour of one alternative over the sum of evidence in favour of all alternatives, is based on axioms and theorems developed in Luce's highly influential work *Individual Choice Behavior* (Luce, 1959).

The basic choice axiom proposes first a finite subset, T, of the universal set U, where for every subset, S, of T, its probability P_S is defined. Where the probability that subset S is selected from subset T is given by $P_T(S)$, the probability that an element of S, called x, is selected $P_S(x)$, is given by the following,

$$P_S(x) = \frac{P_T(x)}{P_T(S)} \qquad (3.10),$$

(*ibid.* p. 7-12).

What this is saying is that the probability of selecting x from S is the probability of selecting x from T, i.e. in general, divided by the probability of selecting S from the larger set T. Put differently this is also describable as the probability of x divided by the sum of all probabilities of the set for which x is a member.

The above axiom refers to situations where, firstly one must assume that only one member of the set T can be selected at a time such that the sum of the probabilities of selection for set T is one. Secondly, one can exclude 'irrelevant' alternatives from the calculation such that the subset, T, is finite with its individual probabilities known. This includes the assumption that none of the probabilities are one or zero. If zero, the element could be excluded from the set T as an irrelevant alternative. If one, then, according to the fact that the sum of the probabilities of members of T is one, the set T would contain only one member.

Luce's choice axiom defines what is meant by choice in this context. In practical terms, the experimenter decides on appropriate responses for the experiment in such a way, generally, as to eliminate any bias on the part of the participant towards one alternative over the others. This may or may not require training or instructions, depending on the experiment and the participant.

The alternatives thus defined, it is generally assumed that other behaviours may be disregarded for the purposes of data collection. This makes sense in, for example, a category learning experiment when the responses of interest may just be the label button pushes. While the participant may exhibit a stronger galvanic skin response to exception members of a category, the evidence that learning has occurred is taken from the button pushes and the feedback which results.

Luce developed the implications of this axiom with regards to its applicability to the behavioural sciences. One of his concerns was with the idea that for various experimental paradigms the interest was in terms of the relationship between individual stimuli and the response probabilities. In this case, one is talking about *conditional* probabilities of a response, given a stimulus. In a simple learning scenario, for example, a participant is first trained to respond 'A' when presented with stimuli x and y, with x and y presented, alone, on separate trials with equal frequency. This training continues until

performance has reached some criterion e.g. the probability of responding A given x or y is close to unity.

In the next phase the participant is shown a compound of x and y, xy. Typically performance generalises to the compound stimulus to some extent. In this case, however, the determination of response probabilities cannot just be a matter of summing the individual response probabilities as these will sum to greater than unity.

Luce suggested that *conditional* probabilities might be described in terms of the sum of any evidence in favour of the selection divided by any evidence in favour of all of the selections. Theorem 3 in his work (*ibid.* p. 27) gives the probability of selecting x from subset T, $P_T(x)$, as follows

$$P_T(x) = \frac{v(x)}{\sum_{y \in T} v(y)} \qquad (3.11).$$

In this case v(x) is assumed to be the value of x according to some ratio scale. The theorem is important because it attempts to characterise choice behaviour as being representable in terms of a numerical scale. It suggests that the probability of selecting an alternative is a function of just the relative weight of evidence. In order to be able to say this, however, one needs to define a scale to describe what constitutes evidence in the first place.

Generalisation indicates that one cannot ignore the influence of previously presented similar stimuli on the probability of a response. The Luce (1963), and Bush *et al.* (1964) model, described in section 3.1.2.1, captures this in terms of its prediction of asymptotic performance, shown in equation 3.9. Here, the asymptotic conditional probabilities consist of the sum of evidence (based on similarity alone), in favour of a particular choice over the sum of the evidence in favour of all of the choices. The scale chosen is one which enables similarity measures to be added such that the alternative with the greatest total of 'similarity' has the highest probability of being selected.

Luce also addressed the issue of calculating response probabilities as a function of the *difference* between the evidence for each alternative as a means of accounting for certain data from psychophysical research. What Luce describes as the Fechnerian position suggests that the ability to discriminate between different extents along a single sensory continuum may be described by some function of the difference between those extents. The *observed* relationship is that the ability to discriminate between two stimuli is a

function of the ratio of those two extents. This captures the observation, for example, that discrimination between the volume of two tones is greater when the two volumes are subjectively low than when they are high, despite the difference on the continuum being the same.

This simple relationship does not apply to all scales. While it may apply for such dimensions as 'loudness' or brightness, other relationships are observed for other dimensions. Discriminablity between different frequencies of tone, for example would appear to follow a cyclic pattern based on harmonics, and temperature discrimination would appear to have two 'ends' with discriminability decreasing as stimuli become colder or hotter (see Shepard, 1987, 1994 for further examples of different modality dependent generalisation relationships).

Luce reconciled this Fechnerian position with the observed data by suggesting that if one first took the logarithms of the continuum values and then used *their* distance from each other in a logistic function, the result was the same as that predicted by the ratio of the two values. So, for the probability of selecting x, as the louder, when presented with a pair of tones x and y, p(x,y),

$$p(x,y) = \frac{1}{1+\frac{v(y)}{v(x)}} = \frac{1}{1+e^{-k(u(x)-u(y))}} \qquad (3.12)$$

where v(x) is the value given to x in terms of a continuum scale of volume, k > 0 and

$$u = \frac{1}{k}\log_e v + a \qquad (3.13).$$

The parameter a is another constant.

The function works by first converting the v values into their logarithms. This scale, when plotted against the ratio scale yields a curve showing decreased difference between log v and log v+n as v increases where n remains constant. As v drops below 1, log v becomes negative approaching minus infinity as v approaches zero. This means that the measure u is more sensitive to small differences between levels of v when v is low than when v is high.

These values obtained, a ratio measurement becomes inappropriate as the u values may be negative. The logistic function in the second part of equation 3.12 is therefore, a means of 'squashing' the difference between the two u values, as they may vary between minus and plus infinity. It also yields the same probabilities as the top function in 3.12 when k=1 and a=0.

This is a fairly important observation in terms of modelling because it highlighted the scope for alternative solutions to the same problem. The approach of Luce was, in a way, to make a 'module' of the choice function. If one's model could not be made, by suitable parameter estimation on the choice function then, due to the axiomatic nature of the choice function used, it may be appropriate to address other aspects of the model while leaving the choice function largely intact.

Luce, in the above example, achieved a fit to the data on pairwise comparison by 'pre-processing' the continuum. If the values representing the evidence for a ratio choice measure are first converted into log values and then compared with a logistic then the differences are captured and the functions are equivalent. Thus Luce may retain the logistic choice function by altering the representation of what constitutes 'evidence' as far as the choice function is concerned.

The ratio approach implies that the decision process is one which involves a finite set of alternatives whose probability of occurrence is neither zero nor one. In a learning experiment one would be assuming naivety on the part of the participants at the beginning but require some guess on each trial. Because the measure of probability from relative evidence is undefined when the sum of the evidence for all alternatives equals zero, this initial condition must be simulated using bias or 'background noise' (Nosofsky *et al.* 1994) constants which give all of the alternatives a baseline amount of evidence. These parameters may or may not be equal but may be any positive real number.

Figure 3.2 shows plots of the ratio and logistic representation of response probabilities, given different levels of difference between a measure of evidence in favour of alternative x and evidence in favour of alternative y (set to zero in this figure). In this figure the logistics are produced using the lower function of equation 3.12, with u(x) and u(y) simply represented by the evidence measures multiplied by a gain or slope parameter k. The ratios are produced using the upper function of equation 3.12. The bias parameters,

b, (equal for each alternative) are simply added to each of v(x) (just x here) and v(y) which being zero means that the ratio is b/x+b.

The bias parameters, if not related to the task, have two functions. The first is to 'enable' a decision when no other evidence is available. The second is dependent on the size of the bias parameters and concerns their role in early learning by such a system. The bias parameters, even if they are all equal, will tend to slow the impact of learning related to the learning rate parameter used as they represent a constant level of error in the system. This can be seen in figure 3.2. With extremely low bias, i.e. a value lower than the learning rate, early learning can be very rapid.

For the logistic approach the decision process is again one which assumes a finite set of alternatives, each with a level of evidence in favour of them. The absolute level of this evidence is unimportant as the function deals with the differences between evidence for alternatives rather than the ratio. Having said this, however, it can be seen from figure 3.2 that the logistic curve is at its steepest when the magnitudes of difference are smallest.

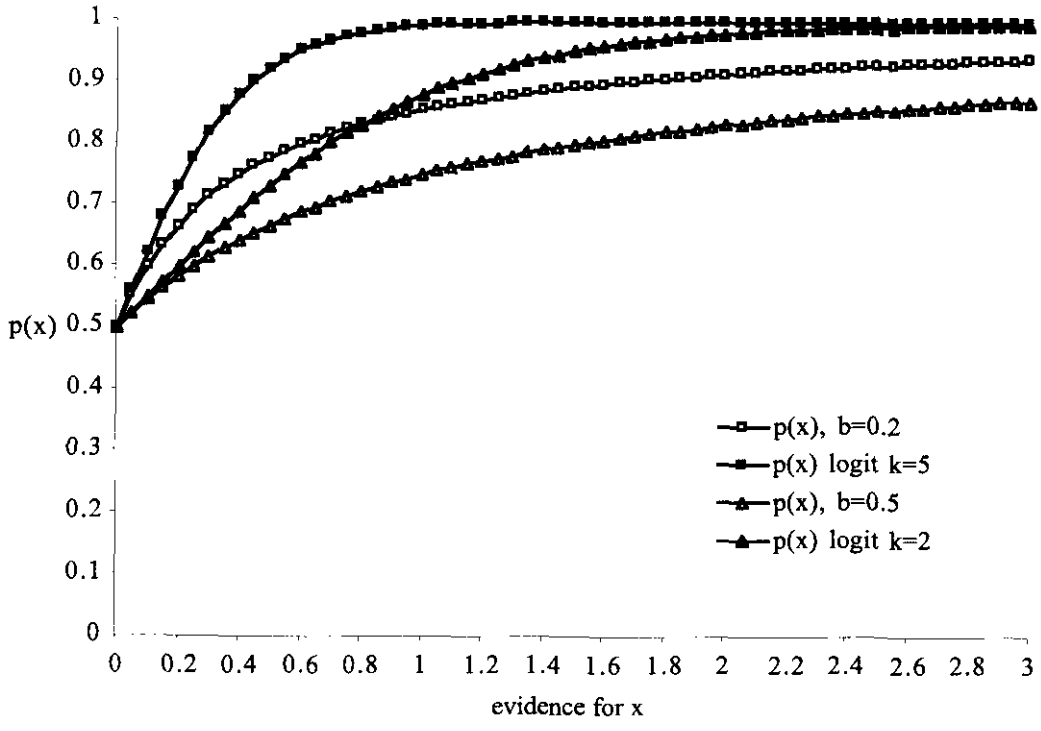

Figure 3.2; Logistic versus ratio estimates of the probability of x given evidence of x as x-axis labels and evidence for y set to zero. For the ratio, two different levels of bias, b, are shown whereas 'gain' or k is altered in the logistic plots.

Experimental evidence suggests that decision-making is compromised by some events and enhanced in others. The logistic displays the same characteristics as a ratio function, if the events which compromise decision making are represented as being able to reduce the difference between levels of evidence in favour of alternatives. Similarly those events which enhance decision making must be represented in a way which leads to increases in the differences between levels of evidence.

This property tends to shift the focus of attention onto the nature of 'evidence' for the choice function i.e. how is its accumulation related to the characteristics of stimuli and their relationship with reinforcement? One corollary of this is that it identifies 'evidence' with concepts such as associative, or habit strength as the 'evidence' may be regarded as the sum of these strengths.

The logistic is more immediately applicable to formulations such as Hull's as measures of response strength could be negative. Hull's approach to describing choice

probabilities was also dependent on some measure of difference between the strengths in favour of alternatives. In this case the combined contributions of excitatory and inhibitory habit strengths for each alternative were subject to a random process described by Hull as 'behavioural oscillation'. This resulted in a 'momentary effective reaction potential' for each alternative, which was made up of the summed response strength plus or minus a value determined by a normally distributed noise process (Hull, 1943, chapters XVII and XVIII). In this case response probabilities would be a function of the difference between the summed response strengths for each alternative and the variance of the noise distribution.

As pointed out by Saul Sternberg, the similarities between the logistic function and the cumulative normal distribution used by Hull to determine the effects of behavioural oscillation make Luce's approach somewhat compatible with Hull's (Sternberg, 1963, p. 30). Hull's habit strengths may simply be viewed as levels of evidence, which are compared in some way by a decision or choice function.

The logistic has also been taken up in various forms as a general means of deriving response probabilities from 'response strength' scales of various types and has proved highly influential in connectionist models. The logistic function, and its often used generalisation when dealing with more than two response classes,

$$p(x) = \frac{e^{v(x)}}{\sum_{y} e^{v(y)}} \qquad (3.14),$$

where v(x) is response strength and y are the alternatives, is particularly useful in generating choice probabilities from response strength scales which include negative strengths.

As discussed above, however, it may commit one to a particular method of representing events in the model such that the logistic may be used as a component of a model fitting experimental data. The next sections on representation and 'selective attention' will further illustrate the role of commitment to model components in determining architectures.

3.1.3 The Rescorla-Wagner learning rule.

3.1.3.1. Three problematic observations from learning research

As stated in the previous section, the stochastic learning models described above may have certain problems when it comes to representing experiments in which the stimuli or the reinforcement schedule change. The alternative approach, which may be described as connectionist in outlook, was to attempt to represent the role of individual stimuli as contributing evidence towards a decision.

The differences between this and Hull's approach would be in the inclusion of specific models of how generalisation affects learning and the use of a relatively straightforward function for deriving response probabilities from relative evidence. The use of individual stimulus representations, with their own potential contributions to evidence, enabled modelling the effects of changing the stimuli or the reinforcement schedule.

Among this type of experiment were a few which could not be adequately represented by versions of the model described above. The following three observations, taken together seemed to indicate that these inadequacies could not be ameliorated by using different ways of representing the stimuli or the response strength.

1. Transfer to compound:
   In this case an animal is first trained on the reinforced pairing of stimulus x and response A, and then tested on a compound of stimulus x and a second stimulus y. Typically it is observed that some of the training on x to A transfers to the xy compound such that p(A) is significantly greater than chance.
2. Transfer to component:
   This experiment is the reverse procedure to the above. The animal is first trained to associate compound xy with rewarded response A. The animal is then tested on stimulus x alone (or y). Here it is typically observed that the training on xy to A transfers to the situation where x or y are presented alone.
3. Blocking:
   This experiment has three parts, beginning with training on the pairing of x and A. This is followed by training with the xy compound and A. The final phase is a test phase in which the component y is presented alone. In this condition the typical observation is

that there is little or no transfer of training from xy to y. As observed by Kamin (1969), with substantial pre-training on the x and A condition, almost no conditioning occurs for y to A. Conditioning of y is said to have been *blocked*.

Blocking is something which has been observed in a variety of experiments with human and animal participants, for a review see Kruschke and Blair (2000), and posed a problem for the learning rules described above. The learning rules described in the previous section imply a certain amount of stimulus independence with respect to learning and reinforcement. The influence of each over the probability of response is a function of its own history of association with the rewarded response and this increments (or decrements) according to the difference between its individual 'control' and the maximum possible for this reinforcement schedule.

Blocking of conditioning cannot be readily predicted by the stochastic and Hullian learning models, regardless, it would appear, of the stimulus representations used. Simply using the individual stimulus components, x and y, with a basic stochastic or Hullian learning model will predict that y will be conditioned to A on the xy compound trials at as rapid a rate as x was conditioned to A on the x to A trials.

The way in which stimulus generalisation models, such as that of Luce (1963) and Bush *et al.* (1964) would cope with such a problem is complicated. Firstly, by the nature of the predictions it makes and, secondly, by the representation of 'similarity' between events with different numbers of 'dimensions' or features. This will be discussed in more detail below in the section concerning stimulus representation.

The solution suggested was that, as far as the increment to associative strength was concerned, the multiple representations acted as a compound such that all active representations might get the same increment on each trial. This increment is determined by the combined associative strength of the stimulus representations.

3.1.3.2. Reinforcement reconsidered: surprise and learning.

The basic intuition of Kamin (1969), shared and formalised by Rescorla and Wagner (1972) was to suggest an elaboration to the nature of the reinforcer. Despite the claims of Estes, mentioned in section 3.1.2 (Estes, 1959, p.404), the way in which reinforcement affected learning *did* require some further examination.

In this case the elaboration was to include the idea of 'surprise' or 'expectation' to the increment. In effect, the model learns nothing about a stimulus' relationship with a reinforced response if that response is already 'predicted' by other stimuli present.

Formally, Rescorla and Wagner (1972) returned to a more Hullian interpretation of theoretically 'unbounded' associative strength, suggesting that for their analyses 'it will generally be sufficient simply to assume that the mapping of [...response strengths...] into magnitude or probability of conditioned responding preserves their ordering' (*ibid.* p. 77). Indexing the associative strength between a stimulus i and a response j as V_{ij}, the increment to associative strength added to V_{ij} after each trial is determined as follows;

$$\Delta V_{ij} = \alpha_i \beta_j \left(\lambda_j - \sum_i V_{ij} \right) \qquad (3.15).$$

The parameters related to the response, j, are the maximum or minimum 'amount' conditionable given a particular reward and response; λ_j, and a 'rate' parameter for that rewarded response; β_j. The rate parameter is assumed to be in the unit interval. The lambdas for each response correspond to a 'target' for the associative strength on each trial. A response may have more than one lambda depending on the context in which the response is produced. As with the Kronecker delta in Luce's identification choice model described above, the lambda represents whether the response was reinforced or correct or not. The 'correctness' of the response depends on the stimulus shown. In a task such as a categorisation or identification task the lambda for a response may be one, when the response is the correct response, and zero when it is the 'wrong' response.

The α_i parameter is a rate which applies to the stimulus component. This was assumed to be unique to that particular stimulus element and was used by Rescorla and Wagner (1972) to approximately indicate 'stimulus salience' capturing the idea that some stimuli are more reinforceable than others regardless of the reinforcement used. The role of these parameters will be discussed in more detail below as alternative solutions to the blocking problem may be produced by varying α_i according to various contingencies during the learning process. Generally the Rescorla –Wagner model held α_i to be a constant property of the stimulus, represented by a value in the unit interval.

The model enables blocking to be predicted, using simple stimulus representations, according to the following reasoning. After x has been conditioned to some level of

responding with respect to A, the presentation of xy, including as it does x, will result in little difference between the sum of associative strengths and the maximum. Regardless of how y is represented in the compound xy, it will only be conditioned to the extent of the difference between the maximum, λ, and the associative strength of however x is represented in the compound xy. This level of conditioning will be further attenuated by the fact that if y is, in principle, conditionable when xy is present, so too is x.

Another general assumption of the model is that the stimulus must actually be present on the trial for its associative strength to contribute to the response. Salience is a term used to denote some aspect of a stimulus which has been observed to be related to reinforceability. Intensity is one example of this where, say, a loud tone is observed to be conditioned at a higher rate than a quiet one. As will be discussed later 'salience' may cover too wide a range of interacting processes to capture adequately using a single parameter.

'Presence', however, will be assumed to be an obvious component of 'salience'. This is generally denoted in the Rescorla-Wagner model by writing the rule in equation 3.15 as

$$\Delta V_{ij} = \begin{cases} \alpha_i \beta_j \left(\lambda_j - \sum_{i \in S} V_{ij} \right) & \text{for } i \in S \\ 0 & \text{for } i \notin S \end{cases} \qquad (3.16)$$

where S is the set of stimulus elements presented on that trial (Sutton & Barto, 1981).

3.1.3.3. The use of Rescorla-Wagner rule

The Rescorla-Wagner rule, despite various shortcomings has been hugely influential and 'provoked' a large amount of research in the area of animal learning. It has also been described as "the primary export of traditional learning theory to other areas of psychology" (Miller, Barnet, & Grahame, 1995, p. 381). One of the main areas for export of the rule is in the use of connectionist models of category learning (Siegal & Allan, 1996).

3.1.3.3.1. Widrow-Hoff and Rescorla-Wagner

Rescorla and Wagner's assertion was that learning took place according to the discrepancy of events from an organism's expectations (Rescorla & Wagner, 1972, p. 75).

Their formalism of this intuition may be regarded as an iterative function of response strengths, stimulus 'presences or absences', and discrepancy. This function results in a reduction in discrepancy from expectations to some minimum determined by a function of the correlation between the stimuli and the responses. Sutton and Barto (1981) identified the Rescorla Wagner rule with the Widrow-Hoff (Widrow & Hoff, 1960), least mean squares (LMS), or delta rule. This rule was developed to solve sets of linear equations by a technique known as 'gradient descent on error'.

In this technique one identifies a set of coefficients by which one set of variables, which may be called the input, may be multiplied to approximate the values of another set, which may be called the output. In this case the coefficient between input i and output j, known as weight w_{ij}, is modified according to the following rule;

$$\Delta w_{ij} = x_i c\left(z_j - \sum_i w_{ij} x_i\right) \qquad (3.17).$$

The z_j is the value of the output variable j, x_i is the value of the input variable i and c is a rate parameter. The rule is also known as a Least Mean Squared rule because, given enough iterations the coefficients or weights will arrive at a solution which minimises the mean of the squared error, where error for each output variable j is the value in parentheses in equation 3.17. Whether this is capable of reducing the error to zero is dependent upon whether the problem is linearly separable.

Linear separability may be described for simple stimulus events using set notation. Figure 3.3 shows an example where the input consists of four 'events', x_1 to x_4, with output events A and B. The diagram represents the task in terms of the members of two sets. One set consists of input events that occur when A is the output and the other contains input events when B is the output.

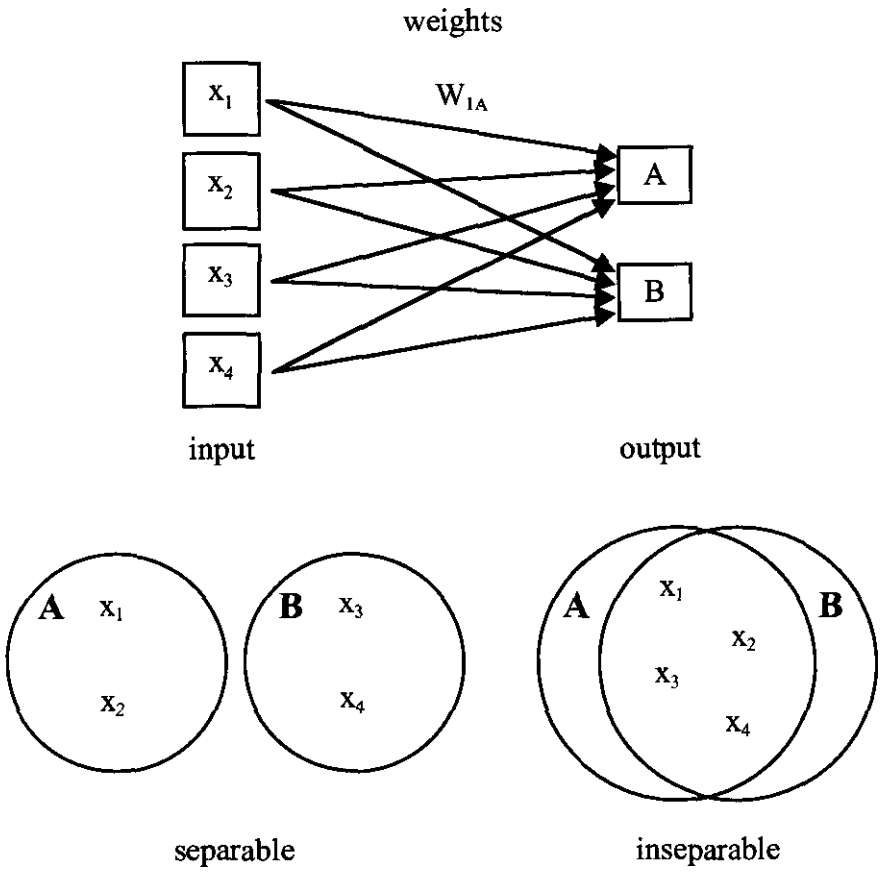

Figure 3.3: Separability of input given output shown using sets. The top part illustrates the inputs and the outputs with assumed weights between each input and each output. The sets beneath are **A**, *inputs which occur when A is the output and* **B**, *inputs which occur when B is the output. In the separable case the intersection between* **A** *and* **B** *contains no members. In the inseparable case some input events occur in the intersection between* **A** *and* **B**.

The task can be said to be linearly separable when,

$$A - B > 0$$

and $$\sum_{x \in (A-B)} p(x \mid A) = 1$$

and $$B - A > 0$$

and $$\sum_{y \in (B-A)} p(y \mid B) = 1$$

where A-B is the set of input events occurring when A is the output minus those members which also occur when B is present, and B-A is the set of inputs which occur when B is the output minus those elements which are also members of set A.

The dependency on 'unique' members of A and B is because members of the intersection may predict either A or B when present. Where the relationship between an input and the outputs is probabilistic, for example x_1 is followed by A three-quarters of the time and B the remaining quarter *and* there are no members of the set A-B present when x_1 is present, the task is not strictly linearly separable. In this case the network or rule may approach a state of its weights where 'error' is minimised but performance will not be perfect.

The conditional probabilities of an input occurring given the output signal are necessary to ensure that there is no 'gap' in the coverage. If these probabilities summed to less than one then there would be times at which no member of the set A-B was present when A was the output.

As pointed out by Sutton and Barto (1981), the Widrow-Hoff, or delta rule is basically identical to the Rescorla-Wagner rule, the main difference being an overt partitioning of the contribution made by a stimulus into an activation and a weight component. Such a partition was implicit in the set theoretic notation used in equation 3.16. In this case the set theoretic notation sufficed as the stimulus representation used was one which would give a value of one to present stimuli and a value of zero to those that were absent.

3.1.3.3.2. Connectionist networks and categorisation

A category, as far as this research is concerned, is a collection of events which, while not necessarily occurring at the same time, are associated with the same response. A category structure may be described as linearly separable if the collection of events associated with one category label minus all those events which are members of other sets, is such that one member occurs with each instantiation of the label.

The same 'rules' apply to category structures as to other input-output mappings if they are learnt, in some way, using a gradient descent technique such as the Widrow-Hoff or Rescorla Wagner rule. These rules made the behaviour of network implementations of

the Rescorla Wagner rule fairly predictable under specified circumstances such as a fixed category structure.

Gluck and Bower (1988a), identifying this logical link between learning theory and categorisation, developed a connectionist implementation of the Rescorla-Wagner rule, coupled with a logistic choice function to see how it would predict certain category learning phenomena. The model used simple stimulus representations but developed the learning and activation functions along lines comparable with those of Sutton and Barto (1981). This meant that the 'activation' of the node was represented in both the response strength generation, and the increment to the associative strength.

The initial model was surprisingly simple, using binary-valued activation for the representation of the presence or absence of stimulus components. It was, however, successful in 'predicting' a phenomenon known as base-rate neglect (Gluck & Bower, 1988a) (which will be discussed below). The representations used in this 'component cue' network were somewhat limiting due to their inability to deal with category structures dependent on configural relationships between stimulus components (which are linearly inseparable for the component cue form of stimulus representation).

One aspect of any category learning task, such as that of Shepard *et al.* (1961), is that it requires adequate representations of the stimuli. Rescorla and Wagner did not commit to using terminology such as 'representation' (Miller, Barnet, & Grahame, 1995). The model is ostensibly concerned with the relationship between particular stimuli and particular responses. The stimuli used and the relationships learnt in the experiments discussed by Rescorla and Wagner were of a fairly simple nature. Stimuli such as tones and lights could be accurately modelled by simply representing each stimulus as a separate 'element' which was either contributing all of its associative strength to a response choice when present, or contributing nothing at all when it was absent.

As research into category learning progressed, network models were developed with alternative methods of representing the stimuli. As will be discussed below, using different representations *can* lead to different model predictions when using the same learning function. In terms of developing a model of category learning, this leads to a question as to what to do when the model does not predict the data. Should one alter the model of representation or alter the learning rule, or do both?

Despite the widespread use of the Rescorla-Wagner rule in models of learning of different types, the rule does have some persistent problems with the modelling of certain observations. In some cases, as the sophistication of category learning models has developed, these shortcomings have been addressed. Somewhat paradoxically though, models incorporating more complex representations and the relationships they imply between stimuli are occasionally unable to continue representing some of the basic data from which the Rescorla-Wagner rule emerged. Some of the difficulties with the basic Rescorla-Wagner model, potential solutions, and the occasional resultant loss of generalisation will be discussed in the next two sections of this review, and returned to later in the thesis.

3.2. Representation of the stimulus

The previous section addressed theories regarding the way in which 'evidence' in favour of a decision to respond in a certain way accumulates according to the conditions of reinforcement associated with that response. It also dealt with functions which seemed able to characterise the way in which evidence in favour of alternative responses may be compared, in order to provide an estimate of different response probabilities. These latter functions indicated that evidence might be compared using either a ratio or a difference measure between their 'levels'. This section addresses the ways used to represent *where* this evidence comes from.

3.2.1. Stimulus representation and learning

It is difficult to say exactly how a stimulus ultimately contributes to the production of a response. All that can be done is to identify what information is required by a given model of the decision process and a given model of the way learning occurs, in order for it to able to represent the observed behaviour.

This seems to be a job of representing which aspects of stimulus variation seem best related to variation in response patterns. All of the learning rules described above attempted to incorporate, in some way, variables representing different properties of the stimuli with respect to learning. For some models, such as the Rescorla-Wagner model, this was in terms of individual salience parameters for each associable stimulus element. The stochastic learning models, such as those proposed by Restle, and Bush and Mosteller attempted to define the properties of the entire stimulus set using a single value. This value was generally derived from some ratio of relevant to irrelevant stimulus components.

Both of these measures were represented in terms of fixed parameters, either pertaining to individual stimuli or the learning rate in general. The idea of linear separability discussed in the previous section, however, suggested that the ability of a learning plus choice model to represent observed data may be substantially improved by altering the representation of the stimuli in the model.

For an associative learning model, the stimulus representation determines how many associative connections are contributing how much to a decision process. It is a

statement about what it is 'about' the stimulus presented which gets associated with a response.

If two events may be perceived as different from one another they may be associated independently with different responses. This independence implies, as far as a learning rule is concerned, that the two events may be represented as having their own sets of associative weights. The 'theorem' described above is somewhat circular, as the only way of demonstrating that two events are discriminable from one another is by getting a participant to respond in one way if they think that event 1 is occurring, and another if they think that event 2 is occurring. The discriminability of the events is then a function of the reliability with which the participant can produce the 'discriminating response'.

3.2.2. Detectors in models of learning

A description, which captures the connectionist idea of an input representation, is that of the 'detector'. The detector may be described as some entity which is maximally responsive to the presence of some particular event. The detector is thus capable of delivering the activation component of the 'evidence' and learning functions if the activation is a measure of the presence of the event.

What the detector is detecting is the modeller's idea of what it is about an event which may be 'captured' or used to inform response probabilities. One may assume at least one detector for each event of interest. The justification for this step may be that the participant can discriminate between the events, therefore separate associative channels capable of developing different response strengths may be involved. The input to these channels must differ depending on which event is actually occurring.

3.2.2.1. Compound stimuli and configural detectors

Another basic finding from learning theory illustrated clearly that whatever these detectors were capable of representing, they must be capable of representing compound stimuli or configural cues as something separate to their components. The basic finding is that training on a set of associations between x and A, y and A, and the compound xy and B is routinely possible for animals. The implication is that animals can discriminate between x and xy to some extent such that a separate association strength for the compound is indicated.

In terms of connectionist modelling this task is basically an analogue of the exclusive OR (XOR) task which was used by Minsky and Papert (1969) to outline the linear inseparability problem for a certain class of network models. The XOR task involves, for a two-input two-output network, the learning of the following mappings 00-A, 01-B,10-B, and 11-A. There is no arrangement of weights possible between x and y and A and B which allows this mapping to be 'learnt' by the network.

The problem generalises beyond compounds of two stimulus elements. Higher dimensionality, or arity, linear inseparability can be described as instances of a 'parity problem'.

A parity problem or a d-bit parity problem (e.g. Bishop, 1995) is one that involves a system trying to reduce mean squared error given a binary input vector with d elements, and output mappings of; A when there is an even number of ones in the vector, and B when there is an odd number of ones. In terms of a learning experiment, the participant is shown stimuli from a set of d possible stimuli. They are rewarded for responding A when there are an even number of stimuli and B when there is an odd number of stimuli; they are not informed that this is a parity problem.

In the Shepard *et al.* (1961) category learning tasks the type VI is an example of a parity problem. In this case inspection of figure 2.1 shows that, when represented as binary sequences, members of category A have an odd number of ones and members of category B have an even number or zero. The type II task is a not a parity problem in three dimensions but *is* in the two dimensions required as a minimum for performance of the task.

As reported in Shepard *et al.* (1961), the re-coding of the type VI problem in terms of a parity problem is quite a rare occurrence. When it occurs it strongly improves performance on reflections and rotations of the type VI structure, which are also parity problems. Parity problems get harder as the dimensionality in which they occur is increased. This may be seen in terms of the comparative difficulty of type II and VI tasks as noted by Shepard *et al.* (1961) and Nosofsky *et al.* (1994). It is also seen at a basic level, in terms of the relative difficulty of condensation and filtration tasks (Kruschke, 1993).

This is a different conceptualisation of the problem in that the condensation task may be represented by binary vectors with the same mapping as that given for the XOR

problem described above. The filtration task has the mappings 00-A, 01-A, 10-B, and 11-B. The filtration task is easier to learn than the condensation task which agrees with the idea that task difficulty may correspond to the number of dimensions at which the problem is a parity problem. Taking zero to be an even number the filtration task is only a parity problem from the point of view of x, whereas the condensation problem is a parity problem at two dimensions.

There are a number of ways of addressing the problem; the simplest first is the configural representation. If a person can do an n bit parity problem then they must have n bit representations of the vector or they have re-coded it such that it can be expressed as a parity problem. Data regarding category learning of stimuli at different arities suggests that firstly, re-coding by participants in terms of a parity problem is rare and, secondly, that the results may be modelled simply using representations with the same arity as the parity problem.

The assumption is that when x and y are presented together they may be represented in terms of an event which is unique to that particular configuration. This event is represented using a separate set of connections or weights to the decision process and a separate activation value (e.g. Wagner & Rescorla, 1972).

Using this configural representation with a choice rule and a learning scheme such as the Rescorla-Wagner rule, the new model is capable of simulating the compound-component discrimination experiment as well as the blocking and the transfer performances described in section 3.1.3.1. This performance is somewhat dependent on the model one uses to determine how the different sources become active.

The simplest approach to this is to suggest that the xy detector is not responsive to the presence of x or y alone, whereas the x and y detectors will activate in the presence of compound xy, as well as when their stimulus is present alone. To suggest that x and y detectors do not activate in the presence of compound xy will remove the ability of the model to generalise performance to the compound xy (problem 1, section 3.1.3.1).

The simple configural approach given will enable transfer to compound, despite the fact that learning to the compound will be blocked by the presence of perfectly relevant x and y cues. Transfer to component will be achieved because x and y detectors are active and learning during the compound trials.

The number of configural detectors for a given stimulus set is, to some extent, dependent on one's interpretation of the detector itself. From a computational point of view, a set of detectors with the same arity as the input vector are all that is required to solve any one-to-one mapping problem. One would also appear to need some means of representing the individual elements to allow transfer to and from compound stimuli. As such the minimum set of detectors required, so far, is one detector for each of the elements and one for each 'total' stimulus presented.

For the Shepard *et al.* (1961) tasks this minimal representation would require that each member of each substitutive feature pair have a detector to itself and each total stimulus have one detector each. With all values of q, r, and s this will result in six 'one-dimensional' detectors and eight, three-dimensional detectors, $2d + 2^d$, where d is the dimensionality of the stimuli or number of pairs of substitutive features.

Using this approach causes problems in terms of representing the differential difficulty of parity problems at different arities. This is revealed by category learning experiments such as that of Shepard *et al.* (1961). This model would predict a difference between a filtration and a condensation task, but would *not* be able to predict the difference in difficulty between the type II and type VI category structures (*ibid.*) shown in figure 2.1 of the last chapter. In any filtration problem the element detectors will gain control over responding quite quickly owing to the fact that they are activated much more frequently than each of the three-dimensional configural detectors is. In any condensation task, regardless of its arity, control will have to be developed by the three-dimensional configural detectors. Their frequency is identical regardless of the task and as such learning of type II and type VI should proceed at an identical rate.

3.2.2.1.1 The basic configural-cue network

There are two solutions to this problem. The first is to apply some measure of similarity as the basis of a detector activation function. As will be described below, in this case generalisation may enable the different difficulties of different arity parity problems to be predicted.

The second way to solve it is to imply that each stimulus set is represented in terms of the activity of a set of detectors corresponding to the powerset of features. For the Shepard *et al.* (1961) tasks, the powerset representation would require detectors for each q,

r, and s feature, each combination of q & r, q & s, and, r & s, features and each configuration of q, r, and s features. This yields a total of 26 detectors which may be grouped into seven sets according to the component pairs represented. As with the binary value stimulus elements this corresponds to q, r, s, qr, qs, rs, and qrs, one detector being active from each set.

Using this representation predicts the difference in difficulty between the type II and type VI structure in terms of the different frequencies of relevant configurations of stimuli. Each of the arity two representations, required for the type II to be learnt (it being inseparable with one-dimensional detectors) is present on one quarter of the trials whereas for the type VI each of the necessary arity three representations are active on only one eighth of the trials.

The configural-cue model was implemented by Gluck and Bower (1988b) using a simple Rescorla-Wagner learning rule and a logistic choice function in order to assess its ability to predict the difficulty of the Shepard *et al.* (1961) category structures. There was no direct representation of similarity in the activation functions with each detector being active only when its particular 'cue' or configuration was present. The model was implemented with a single output node using values for lambda of 1 for one category and −1 for the other. While it predicted the relative difficulty of types I, II, and VI, it predicted the overall order to be different to that observed. The configural-cue model predicted that for early blocks III and IV were learnt with fewer errors than types II and V. Later in the simulation though II and IV change order such that the final order is I, III, II, IV, V, VI.

Nosofsky *et al.* (1994), in their partial replication of Shepard *et al.* (1961), tested the configural-cue model using a slightly different learning rule and a ratio-based choice function in order to establish a quantitative fit to the learning data using parameter optimisation. They revealed a best fit which resulted in an ordering pattern and crossover qualitatively identical to that found by Gluck and Bower (1988b).

The reason why the configural-cue model does not predict the correct order of difficulty will be discussed in more detail below and in the next chapter. The failure of the basic configural-cue model to reproduce the order of difficulty observed by Shepard *et al.* (1961) and Nosofsky *et al.* (1994), however, indicates only that this type of representation with the Rescorla-Wagner learning rule and a logistic or ratio choice function is

inadequate. Attempts to get a model using a configural-cue representation to reproduce this order have consequently focused on altering the way in which detectors become associated with responses. These attempts involve adding extra weights controlled by their own learning rules to the model as an augmentation of the basic associative rules. These modifications will be discussed in more detail in the third section of this chapter when models of selective attention are discussed. Before that, the next sub-section deals with other models of stimulus representation.

3.2.2.2. Similarity, generalisation, and the psychophysical model

As discussed in the previous sections, generalisation of response control between similar stimuli is a central part of any theory of learning. In terms of the activation of a detector or set of detectors representing stimuli, one's model of generalisation may determine, to a great extent, the activation function for each detector.

The use of contingency tables to represent confusability originated with psychophysical research into such areas as discrimination and identification performance. In this domain it is traditionally known as a confusion matrix. In the table one may represent the different responses by their own row and the stimuli by their own column. In each cell the number of times a participant made a particular response when presented with a particular stimuli is recorded. These numbers may be converted into joint probabilities p(stimulus, response) by dividing the number in the cell by the total number of responses made. Generally only one response is made per trial, where only one stimulus is presented, in the case of identification. For discrimination tasks two stimuli may be presented with the participant required to respond with one behaviour when they think that, say, tone stimulus 1 is the louder and respond with the other when they think stimulus 2 is the louder.

Obviously, only one cell in each row contains 'correct' responses. The remainder are generally known as 'confusions'. While the error could be random, the analysis of these results, on an enormous number of occasions with a wide variety of sensory continua has revealed that this is extremely unlikely (see Shepard, 1987 for a brief review). The probability of responding that stimulus x has been presented in an identification task when it was, in fact, y decreases as a function of the 'distance' between x and y on the continuum.

In his review of detection and recognition theory in 1963, Luce describes the development of the similarity measure used in the model described in section 3.1.2.1. The basic property of the similarity measure is that if one has a continuum representation of a particular stimulus' level or intensity, it is demonstrated that the probability of confusing this stimulus for another, with a different level, is a function of a constant raised to the negative power of the difference between the two levels on the continuum scale.

In order to determine the choice probabilities of responding that x has been presented rather than y, when a stimulus i has been presented between them on the continuum scale, one simply compares the similarity of the stimuli to the two 'detectors'. Figure 3.4 illustrates graphically the relationship between distance on a scale, similarity, and the kind of response probabilities observed.

In this case, two stimuli have been presented at level 1 and level 5 on the continuum. The participant is told the appropriate response for each stimulus. Testing with stimuli in between 1 and 5 reveals a sigmoid pattern of response probabilities as shown with a point of maximum uncertainty (where $p(1|i) = p(5|i) = 0.5$) halfway between the two levels.

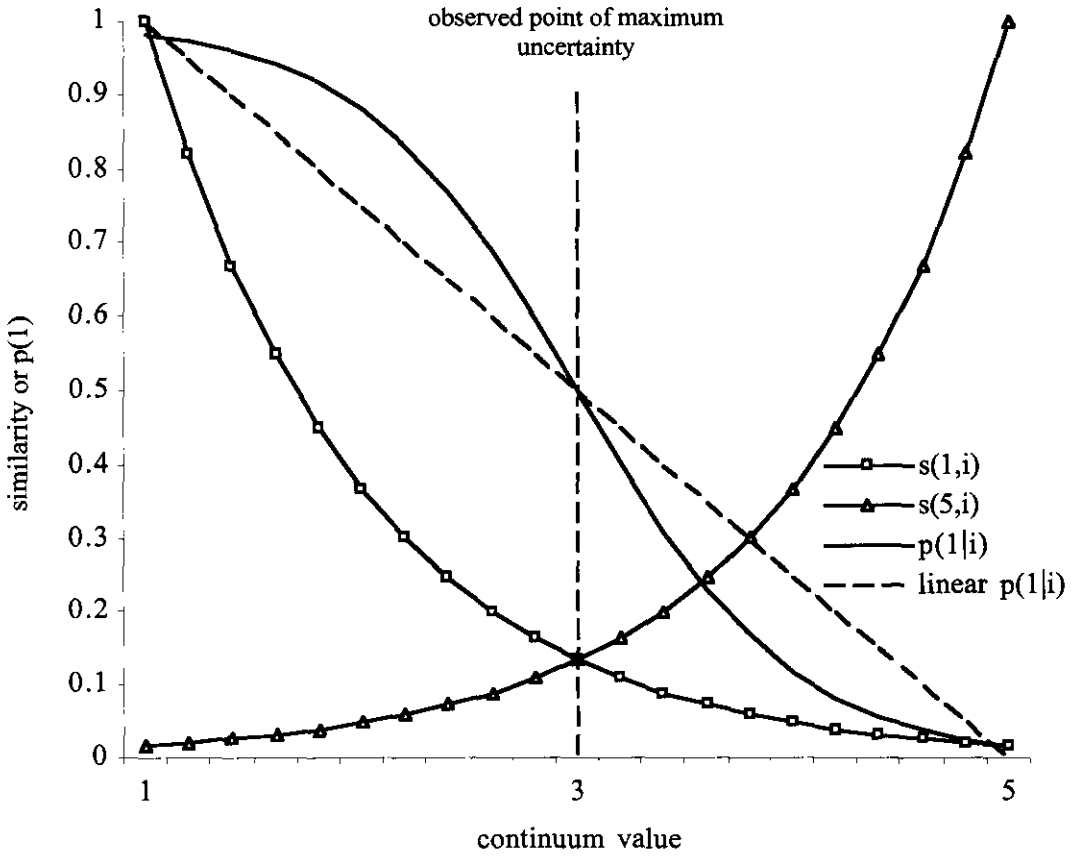

Figure 3.4: Relationship between distance along a continuum of a stimulus i, its similarity to a stimulus located at 1, s(1,i) and to another at 5, s(5,i). The probability estimate, p(1| i) is calculated from the ratio of similarity to 1 over the sum of similarities to 1 and 5. A 'linear' estimate of conditional probability is also shown. Both probability measures share the point of maximum uncertainty at halfway between the two stimuli.

The simple linear distance function;

$$p(1\,|\,i) = 1 - \frac{|1 - l_i|}{|1 - l_i| + |5 - l_i|} \qquad (3.18)$$

where l_i is the level of i on the scale will yield the diagonal dashed line shown in figure 3.4 which is at odds with the observed distribution. The ratio of similarities, however, produces something approaching the observed distribution.

$$p(1\,|\,i) = \frac{s_{1i}}{s_{1i} + s_{5i}} \qquad (3.19)$$

Here s_{1i} is the similarity of the input, 1, to i and, assuming that the similarities are symmetrical, the similarity of i to 1. The similarity of two stimuli i and j on a continuum, x, for use in the above probability estimate, is given by,

$$s_{ij} = e^{-c|x_i - x_j|} \qquad (3.20)$$

where x_i and x_j are the levels of i and j on continuum x and c is a scaling parameter, $c \geq 0$. The scaling parameter is at the heart of the relationship between the similarity measure and the response probabilities. The scaling parameter here is generally constant across the scale. For figure 3.4, c is set to unity. Increasing the parameter will increase the gradient on the similarity curves, the result of this being a steepening of the sigmoid choice probability curve around the mid-point, with it approaching asymptote on either side at a more rapid rate.

For this model it is the scaling parameter that has to be estimated from experimental data by minimising the discrepancy between the choice curve predicted and that observed. This parameter determines what size of difference on the continuum, with its arbitrary scale, may represent a unit of 'psychological distance' or psychological difference. For figure 3.4, for example, if the observed curve between 1 and 5 units on the scale of interest was steeper around the mid-point than that shown, the implication would be that one unit on the measured scale corresponded to more than one unit in psychological space. This would mean that c would have to be increased to 'scale' the input.

This similarity measure is the same as that used by Luce (1963) in his identification choice model and Bush *et al.* (1964) in their more general treatment of the approach, its value is represented in equations 3.8 and 3.9, given in section 3.1.2.1, as $\eta(i,j)$.

3.2.2.3. Exemplar representations, similarity, and categorisation

As discussed briefly in the previous chapter, Shepard *et al.* (1961) considered that the relative difficulty of category structures might be modelled in terms of how similar the members of different categories are to each other. The stimuli used for categorisation experiments were generally ones which may have been described in terms of collections of features rather than differences along different continua.

In this case there is a difference in the kind of scales used to measure the distance between two objects. For the experiments carried out in psychophysics, the stimuli would vary along continuously differentiable quantitative scales. These scales would typically be

ratio scales. In the case of categorisation experiments and many of those in animal learning, the differences between stimuli tended to be of a substitutive nature with respect to their features or components. A feature would either be present or absent, or two features such as a square and a triangle may form a mutually exclusive pair, one of which is always present in a presented set of features or object.

3.2.2.3.1. The context model

In 1978 Medin and Schaffer proposed a new model of category judgement and learning which was based in the idea that people made categorisation decisions based on the retrieval of information from 'exemplars' stored in association with the category 'label' or rewarded response. The exemplar is an individual instance of an object or total stimulus. The representation of the stimulus here is in the form of one 'node' or detector per stimulus instance. The learning rule proposed for this model will be discussed later but, concentrating on the detector, the activation function for this representation is of interest.

Evidence contributing to a category decision would be retrieved from each stored exemplar by a particular input, according to the similarity of the stored representation to that input. A 'multiplicative' similarity rule was proposed based on the intuition that the presence or absence of some features would be more diagnostic of category membership than others (Medin & Schaffer, 1978).

The multiplicative rule represented the presence of a feature with a one and its absence, or alternate form, with another number whose value was in the unit interval. The multiplicative rule meant that if the presence of a feature was a necessary condition of class membership, then its absence in a particular stimulus could be represented by a zero. Regardless of the number of other valid features present in the object, once their product was multiplied by the zero, the similarity, and thus the contribution, of the stored exemplar to the category judgement would be zero.

The determination of the value of such parameters in a given experiment will be discussed below but, the similarity of an object i to a stored exemplar j, s_{ij} with number of features n_j is determined as follows;

$$s_{ij} = \prod_{k=1}^{n_j} f_k \qquad (3.21).$$

The value f_k is the similarity value for feature j where $0 \leq f_k \leq 1$ if that feature is absent or different in the object and 1 if it is the same. In the case where all f_k are equal to a single value s when the feature is absent or different, the similarity may be described in terms of the number of feature values for the exemplar j which are different in the object i, d_{ij} , as follows,

$$s_{ij} = s^{d_{ij}} \qquad (3.22).$$

In order to predict the probability of a particular category decision, given a particular test stimulus, Medin and Schaffer (1978) used the sum of the similarities of the input to the stored members of that category divided by the sum of the similarity to all stored exemplars. So the probability of selecting category A from categories A and B, given test stimulus i is,

$$p(A|i) = \frac{\sum_{a \in A} s_{ia}}{\sum_{a \in A} s_{ia} + \sum_{b \in B} s_{ib}} \qquad (3.23).$$

To make predictions regarding experimental data on assignment of transfer stimuli, the f parameters must be estimated from the training data. Medin and Schaffer did this by optimising parameters to achieve a close quantitative fit of predicted error to the observed error frequencies for the experiment.

Medin and Schaffer compared the performance of the context model with that of component cue models and also a prototype model on tasks involving performance on new objects similar to members of training sets whose assignments had previously been learnt. The category structures were 'ill-defined' or not linearly separable using component cue representations so the superiority of performance by the exemplar model is unsurprising in this respect.

The basic idea of prototype models (e.g. Rosch & Mervis 1975) is to assume that category learning establishes a category prototype, or average member, for each category based on the frequency of occurrence of different features within each category. The probability of category membership for a particular stimulus is calculated in terms of its 'prototypicality' which may be described in terms of the similarity of the input to the category prototype.

These mathematical models appear to have little capacity to generalise beyond some 'family resemblance' structures, where membership may actually be a function of proximity to the prototype. The type IV category structure of Shepard *et al.* (1961) shown in figure 2.1 is an example of a family resemblance structure in three dimensions. The prototype, or central members may be described in terms of having feature values of all ones or zeros. The more ones a stimulus has, the closer it is to the prototype and, the argument goes, the more readily it will be assigned to that category. Other structures from the Shepard *et al.* (1961) experiment cannot be readily predicted using a prototype approach (Estes 1994, p. 51-54). The type II and type VI structures, for example, have their prototypes 'located' in the same place; i.e. they are identical for each category.

The experiment designed to contrast the performance of the context model from the prototype model involved a family resemblance structure in four dimensions. There were sixteen stimuli, nine of which were presented during training. These, significantly, surrounded (in terms of number of different features) but did not include the prototypical member for one of the categories.

The prototype model would predict that when that prototype was presented as a transfer input it should lead to the appropriate category response much more frequently than would a member of the training set with just one feature different to the prototype. This is because the prototype is effectively already represented as the average training set member, with proximity to that member determining level of 'output' from the prototype.

The context model would predict that the previously presented exemplar would be classified more accurately than the prototype because they would be closer to the other members of the category than the prototype, which would have to be at least one feature different to all other members. The results contradicted the predictions of the prototype model, but were predicted to some extent by the context model (Medin & Schaffer, 1978, Nosofsky, Kruschke, & McKinley, 1992).

3.2.2.3.2. The identification-categorisation relationship: the mapping hypothesis.

The resemblance between Medin and Schaffer's (1978) conception of the determination of response probabilities and that presented in Shepard's model of stimulus generalisation (Shepard, 1957) and Luce's identification choice model (1963) was examined in detail by Robert Nosofsky (Nosofsky, 1984, 1986). Nosofsky related the

identification model with the context model of categorisation via *the mapping hypothesis*. The hypothesis originated in Shepard *et al.* (1961) as a way of describing the difficulty of the categorisation tasks in terms of the similarity of members within categories and the dissimilarity of members of different categories.

This involves the assumption that an identification learning experiment could be conducted with the stimulus set. The frequency of each stimulus-identifying response could be presented in an identification-confusion matrix, with one row for each stimulus presented and one column for each identifying response. Having done this, one could estimate the difficulty of a categorisation task, where stimuli were grouped into two equal sets. According to the mapping hypothesis, the probability of assigning each stimulus to the correct category would be a function of how likely the participant was to confuse the stimuli, in the identification experiment, for stimuli which were, in this case, members of the same category (Nosofsky, 1984, p.105).

A formal representation of the mapping hypothesis is given by Bush *et al.* (1964) in the context of identification experiments where, when presented with a stimulus k during training the participant was, with probability represented by $\Pi(j|k)$, given feedback saying that the stimulus was, in fact, j. As with the identification choice function in equation 3.9, this function describes the expected conditional probability of producing response j when presented with stimulus i as the number of trials, n, approaches infinity (from Bush *et al.* 1964, p. 212),

$$\lim_{n\to\infty} E\left[p_n(j|i)\right] = \frac{\theta(j)\sum_{k=1}^{m} P(k)\Pi(j|k)\eta(i,k)}{\sum_{h=1}^{m}\theta(h)\sum_{k=1}^{m} P(k)\Pi(h|k)\eta(i,k)} \qquad (3.24).$$

$P(k)$ is the probability of stimulus k, $\theta(j)$ is the response-determined learning rate of response j, and, as above, $\eta(i,k)$ is the similarity of i to k. If one assumes the learning rates to be equal for all responses, and the probability of each stimulus to be equal, the equation reduces to,

$$\lim_{n\to\infty} E\left[p_n(j|i)\right] = \frac{\sum_{k=1}^{m}\Pi(j|k)\eta(i,k)}{\sum_{h=1}^{m}\sum_{k=1}^{m}\Pi(h|k)\eta(i,k)} \qquad (3.25).$$

The numerator is the sum of the similarities of i to all m k multiplied by the probabilities that when k was presented, j was the rewarded response. The denominator is the sum for all m responses, h, of the similarities of i to all m k multiplied by the probability that k was assigned to response h.

In a category learning experiment such as the Shepard *et al.* (1961) paradigm, the probability of rewarded response will be one or zero. With only two responses A and B, four of the stimuli, k, have $\Pi(A|\, k) = 1$ and $\Pi(B|\, k) = 0$, the remaining four have $\Pi(A|\, k) = 0$ and $\Pi(B|\, k) = 1$. This reduces the function to that effectively used by Medin and Schaffer (1978) given in equation 3.23, and described by Nosofsky (1984) as a bias-free version of Luce's (1963) choice model for stimulus identification.

3.2.2.3.3. Identification, categorisation and similarity.

It is important to note that despite the similarity of equations 3.25 and 3.23, they are actually referring to different things. The Luce (1963) and Bush *et al.* (1964) equations both refer to asymptotic response probabilities for both previously presented and transfer stimuli. Equation 3.23 refers to only the response probabilities for transfer stimuli presented after asymptotic training on a set of similar, but non-identical, stimuli. It is also applied as a learning rate estimate as it was, in a way, by Shepard *et al.* (1961) and to the same structures by Nosofsky (1984). The six category structures are all 'learnable' typically to a criterion of no errors across 32 consecutive trials.

This suggests, as an implication of the identification choice rule, that the similarity between the objects presented is, effectively zero if measured in terms of their psychophysical dimensionality. The similarity/ generalisation referred to in category learning tasks, where the stimuli consist of substitutive (and one assumes perfectly discriminable) feature collections, is of a somewhat different nature to generalisation, as used in identification and recognition paradigms. The latter form of generalisation is related to some assumption of limitation in the capacity of sensory systems to resolve differences of less than a certain absolute magnitude. The effect of this cannot be removed by learning.

In common, the two forms may both influence the rate at which asymptotic response probabilities are reached. As suggested above the only asymptotic probabilities that featural object similarity will control is that of transfer stimuli. This is, of course,

unless the task is not linearly separable at the arity of the exemplars themselves. For example, some stimuli are sometimes members of A and the rest of the time members of B, such that these stimuli are identical to stimuli in both categories. Whatever the nature of this generalisation, it may be removed by learning and, as such does not represent a limit to the ability to categorise the stimuli.

Despite this important difference, the nature of the generalisation function for objects such as those used in Shepard *et al.* (1961) may be described in similar ways. Nosofsky (1984) considered the multiplicative similarity rule of Medin and Schaffer (1978) in relation to the negative exponential used in psychophysical models.

In the identification choice model and in multidimensional scaling theory, similarity is regarded as a decreasing function of 'psychological' distance. Such a model, when applied to multidimensional stimuli, involves the stimuli being located in points in a psychological space with as many dimensions as the objects have. Their similarity is then a function of the sum of the distances between them on all dimensions.

The distance metric used in multidimensional scaling (MDS) theory is known as a Minkowski-r metric and takes the following general form where d_{ij} is the distance between stimuli i and j,

$$d_{ij} = \left(\sum_x |i_x - j_x|^r \right)^{\frac{1}{r}} \qquad (3.26).$$

Each x is a dimension with i_x and j_x representing the scaled (using identification and MDS) values of stimuli i and j on that dimension. The value of r, when 2, renders the 'psychological distance' a function of the Euclidean distance in n_x dimensional space. It has been found that Euclidean distance (r=2), only seems to provide adequate confusion estimates for stimuli composed of what may be described as *integral* dimensions.

Integral dimensions are those which appear to 'combine' into unanalysable wholes when together in the same stimulus (Nosofsky, 1984). An example appears to be the dimensions of hue, brightness, and saturation with respect to colours (Nosofsky & Palmeri, 1996). These may be contrasted with *separable* dimensions which appear to remain psychologically distinct when combined in a stimulus. The stimuli used in categorisation and learning experiments, particularly those described so far, tend to be of this type.

For separable dimensions, a Minkowski-r metric of 1, also known as the 'city-block' metric, or one which describes the 'Hamming distance' (Gluck, 1991) is generally used. Whereas the Euclidean metric is a straight-line distance between points, the city block metric may be described as the distance between two points when one can only travel along one dimension at a time. It is important to note that the r metric is an estimated parameter. It should be emphasised that an r metric of two enables better fits to data for stimuli composed of dimensions described as integral, whereas r=1 seems to fit stimuli describable as composed of separable dimensions.

Some justification from categorisation research for the distinction between analysable and non-analysable, applied to dimensions for which r values of 1 and two are appropriate, emerges when concepts of selective attention are considered. This will be discussed in more detail in the following section on selective attention.

Nosofsky (1984) pointed out that one could provide a certain amount of psychological justification for Medin and Schaffer's (1978) multiplicative similarity rule by identifying the relationship between it and the distance measure. This comes from the assumption that distance influences decision processes according to the exponential decay function used in Luce's (1963) model. In this case the f_d values for the multiplicative similarity function in equation 3.21 are each determined using the exponential decay similarity function given in equation 3.20. Relative to equation 3.20, f_x is the component similarity of two objects relative to component x and d_{ij}^x is the difference between two objects i and j on that component, either zero (the same) or one (different),

$$f_x = e^{-cd_{ij}^x} \qquad (3.27).$$

The scaling parameter c controls 'effect' of the difference, the higher the value of c, the lower f will be when d=1. The similarity of two objects, i and j, may thus be expressed in the following ways;

$$s_{ij} = e^{-c\sum_x d_{ij}^x} = \prod_x e^{-cd_{ij}^x} \qquad (3.28),$$

Nosofsky (1984, p. 107), where d_{ij}^x denotes the difference between i and j on component or dimension x.

3.2.2.3.4. Application of the exemplar similarity approach to Shepard et al. (1961)

As discussed in the previous chapter, Shepard *et al.* (1961) attempted to predict the relative difficulty of the six category structures using a stimulus generalisation approach equivalent to the above. As also pointed out, the approach failed to predict the observed relative difficulties.

Table 3.1 shows the similarity of each of the eight objects (from Shepard *et al.*, 1961) to each other, assuming that each dimension contributes an equal amount to the similarity measure. When one feature is different, the similarity between two objects is, consequently, s. According to the multiplicative similarity rule, when two features are different the similarity will be s^2, and with three different it will be s^3. This is equivalent to saying that c, in equations 3.27 and 3.28, is equal for all dimensions or feature pairs. Identical stimuli have a similarity of one.

The table enables 'prediction' of the confusion matrix for these eight stimuli where the probability of confusing 000 for 001, for example, may be estimated by dividing the value in the cell (000,001) (row, column), by the sum of the values in the row 000.

The mapping hypothesis, as described above, suggests that the difficulty of a category structure may be described in terms of the probability of confusing a stimulus from category B with other members of category B. In the table this may be evaluated for the type I structure by simply dividing, for each stimulus the sum of the similarities in the bold-headed columns by the sum of all the similarities in that row.

	000	**001**	**010**	**011**	100	101	110	111
000	1	s	s	s^2	s	s^2	s^2	s^3
001	s	1	s^2	s	s^2	s	s^3	s^2
010	s	s^2	1	s	s^2	s^3	s	s^2
011	s^2	s	s	1	s^3	s^2	s^2	s
100	s	s^2	s^2	s^3	1	s	s	s^2
101	s^2	s	s^3	s^2	s	1	s^2	s
110	s^2	s^3	s	s^2	s	s^2	1	s
111	s^3	s^2	s^2	s	s^2	s	s	1

Table 3.1: Representation of the similarity matrix for the eight stimuli from the Shepard *et al.* (1961) task with a binary number representation of each heading each column and row (see figure 2.1). Bold type shows members of category B in the type I category structure.

Assuming a value for c of one, and the distance represented by one feature difference to also be one, the appropriate ratio for each stimulus and each category structure can be calculated. Table 3.2 shows the ratios, as determined using equation 3.23, as they apply to the four members, in each structure, of category B.

As can be seen, the rank ordering of the average conditional probability is I, IV, III, V, II, and VI. This is the same order as that reported by Shepard *et al.* (1961) for their estimation, and subsequently by Nosofsky (1984). As can also be seen there are differences in the conditional probabilities predicted for different members for the types III, IV and V structures. Central members of the category (see figure 2.2) have the highest probabilities followed by peripheral members, followed, in the case of structure V, by exception members. Manipulating the value of c for the similarity functions does not have any effect on the overall ordering of the data, provided, of course, that $c>0$.

	category structure					
member of B	I	II	III	IV	V	VI
1	0.73106	0.60678	0.60678	0.64020	0.51592	0.54934
2	0.73106	0.60678	0.60678	0.64020	0.69763	0.54934
3	0.73106	0.60678	0.73106	0.64020	0.64020	0.54934
4	0.73106	0.60678	0.73106	0.82191	0.64020	0.54934
average	**0.73106**	**0.60678**	**0.66892**	**0.68563**	**0.62349**	**0.54934**

Table 3.2: Probability of response B, given each of the members of category B, for each of the six category structures according to equation 3.23 and the relative similarity measures given in table 3.1. The bottom row shows the average probability.

3.2.2.3.5 Learning in exemplar models: the exemplar network

The choice of learning rule for exemplar models of identification and categorisation, is a matter which highlights the difference between these models and the identification choice model of Luce (1963) and Bush *et al.* (1964). As discussed above, the mapping hypothesis relating categorisation and identification is concerned with a different 'type' of similarity and generalisation to that represented in psychophysical models. In the case of categorisation models, the ratios derived from an identification confusion matrix may generally be applied to learning rates rather than asymptotic response probabilities.

Early variants of the context model (Medin & Schaffer, 1978, Estes, Campbell, Hatsopoulos, & Hurwitz,1989) were tested using a learning model which simply involved adding an exemplar to a category each time one was presented and associated with the category label. Each time a stimulus is presented, the response probability for a category, A, may be determined by the ratio of similarities of the stimulus to exemplars previously 'placed' in category A, divided by the similarity of the stimulus to all exemplars.

As discussed in section 3.1.2.2, the learning curve produced by this approach is one that tends to approach the asymptotic values defined by the Luce model. There is no leaning rate parameter for this model as the change in probability is determined by the

effect of the addition of one exemplar to a particular category. As also described in 3.1.2.2. the rate, in this case, will be controlled by the addition of bias or background noise parameters to the choice ratio function. These control the rate at which the model's choice probabilities move away from chance, in terms of the rate at which the addition of new exemplars marginalises the influence of this bias parameter on the choice function.

Evaluating the exemplar model in relation to component cue networks, Estes *et al.* (1989) suggested that the problem of these low asymptotic response probabilities, described as the 'overlap problem' (Rudy & Wagner, 1975), may be eliminated by the use of associative weights on each exemplar. The alternative is to suggest that the similarity parameters, or the gradient on the decay function, alter across the course of learning such that generalisation decreases as learning progresses. This leaves the problem of the exemplar model being unable to account for transfer performance at asymptote as, by this point, it is no longer able to generalise (Estes, 1994, p. 76).

As will be discussed below, some models based on the exemplar representation effectively do both, i.e. alter similarity measures and use adaptive weights on each exemplar detector (e.g. Kruschke, 1992). These models incorporate assumptions regarding selective attention to particular stimulus dimensions and will be described in detail in the next section.

In the simple exemplar network model, however, rather than adding an exemplar each time a stimulus is presented, the exemplar is conceptualised as a detector with an associative weight connected to each of the category decisions. This weight simply gets incremented by a variant of the Rescorla-Wagner learning rule according to the activation of the detector (related to the similarity of the stored exemplar to the stimulus presented), the 'teacher signal' at each category label node, and a constant learning rate.

Variants of this model were developed and tested by Kruschke (1992), Nosofsky, Kruschke, & McKinley (1992), and Estes (1994) on various category learning experiments where its performance was generally superior to that of the context model of Medin and Schaffer (1978). In particular, Nosofsky *et al.* (1992) applied the model to a replication and extension of the inverse prototype experiment described in Medin and Schaffer (1978). The extension of the experiment involved tracking the inverse prototype effect at regular intervals across the training phase where transfer stimuli were presented after every four

blocks of training data. The model performed better than the context model on this extension, a superiority which may be attributed to the use of the Rescorla-Wagner learning rule (Nosofsky *et al.* 1992).

The reason for this success is that the Rescorla-Wagner rule enables exemplar based detectors to develop inhibitory associative weights with respect to the categories they are not members of. These inhibitory weights may, as learning progresses, cancel out the excitatory contribution made by generalisation-activated exemplars belonging to a different category.

The learning rate may still be predicted as being proportional to the intra-category similarities and inter-category differences of stimuli. In this case it determines the rate at which an exemplar's associative weights are modified in favour of its category, compared to the rate at which it is receiving the opposite pattern of signals due to its activation by generalisation from a member of the alternative category.

At this point, however, the difference between the effects of generalisation in categorisation models such as the exemplar network and that represented by models such as Luce's identification choice model (1963) (as discussed in section 3.2.2.3.3) becomes signifcant. As discussed, the generalisation in the Luce model relates, at asymptotic levels, to some limitation of sensory capacity which does not appear to be compensatable by learning. In categorisation tasks the measure is one of generalisation of response. For many of the stimuli employed in category learning tasks, the probability of actually confusing one stimulus for another is likely to be vanishingly small. The measure used is based on multidimensional scaling, where the participant is asked to state which of two stimuli a third probe stimulus is more similar to.

For familiar, readily discrimininable stimuli, when learning has proceeded to asymptotic levels, generalisation is not observable. This either indicates that its effect is 'concealed' by the interaction of positive and negative associative weights (as suggested by the exemplar network approach), or that it is simply not occurring at all. In solving the overlap problem using the Rescorla-Wagner learning rule, the exemplar network loses the capacity to represent the sensory capacity limitation which results in psychophysical measures of confusability.

3.2.2.4. Comparison of the exemplar and configural-cue forms of representation

Comparison of exemplar models with other approaches is a complex affair due to the different assumptions each model usually makes regarding the representation of the stimulus. For example, in the case of the configural-cue model, as described above, the similarity of two stimuli is 'manifested' at the level of the combined contributions from multiple representations or detectors. In the exemplar model, the activation of the detectors themselves is a function of their similarity to the input.

The similarity functions of the two approaches may be contrasted in terms of the different accounts given by each model of the process of generalisation. In the exemplar model, generalisation occurs due to partial activation of the representations of whole stimuli due to their similarity to the one presented. The decision process is compromised by generalisation because members of different categories will make, albeit attenuated, contributions to each category decision in favour of the category to which they belong.

In the case of the configural-cue model, generalisation takes place not because of some discrimination failure, but due to the use, in a decision process, of cues and cue configurations which are not unique to the stimuli presented. Where these cues and cue configurations are shared by stimuli from other categories, the decision process may be similarly compromised.

The similarity function of the configural-cue model, with respect to the initial similarity of stimuli, may be calculated in terms of the number of features and feature configurations each stimuli has in common. Assuming a situation where, for example, a configural-cue network has been trained to asymptote on a single three-dimensional stimulus, one could say that equal associative strength would accrue to each of the seven cue and cue configuration detectors which represent the stimulus. On being shown a second stimulus with a given number of features, d, different from the trained example, the similarity measure, in terms of response strength which may be delivered to the decision process may be evaluated as follows;

$$s_{ij} = \frac{2^{(n-d)} - 1}{2^n - 1} \qquad (3.29),$$

where n is the number of dimensions in i, and d is the number of different dimensional values in j. The value 2^n-1 always gives the number of 'spaces' required by a configural

cue model to represent a stimulus with n dimensions and also the number of detectors which will be active, one from each space, when a stimulus with n dimensions is presented.

Like the exemplar similarity function, this function also provides exponential decay of similarity as the number of features different, or the Hamming distance, increases (Gluck, 1991, Shanks & Gluck, 1994). One difference, however, is that when a stimulus has no features in common with one represented by a configural cue network, the similarity is zero. Another difference is that the gradient of the similarity function changes according to the number of features the object has. The gradient is inversely proportional to the number of features.

Applying this function to the confusion matrix approach and the mapping hypothesis for the Shepard *et al.* (1961) stimuli yields the similarity measures shown in table 3.3.

	000	**001**	**010**	**011**	100	101	110	111
000	1	3/7	3/7	1/7	3/7	1/7	1/7	0
001	3/7	1	1/7	3/7	1/7	3/7	0	1/7
010	3/7	1/7	1	3/7	1/7	0	3/7	1/7
011	1/7	3/7	3/7	1	0	1/7	1/7	3/7
100	3/7	1/7	1/7	0	1	3/7	3/7	1/7
101	1/7	3/7	0	1/7	3/7	1	1/7	3/7
110	1/7	0	3/7	1/7	3/7	1/7	1	3/7
111	0	1/7	1/7	3/7	1/7	3/7	3/7	1

Table 3.3: Representation of the similarity matrix for the eight stimuli from the Shepard *et al.* (1961) task with a binary number representation of each heading each column and row (see figure 2.1). These measures are produced using equation 3.29 as a similarity function for the configural-cue network. As in table 3.1, bold type shows members of category B in the type I category structure.

These values can be used in the same way as for the exemplar model, above, to calculate the ratios for each stimulus and each category structures. This results in similar patterns of individual ratios and average choice ratios for the structure as indicated by table 3.4.

	category structure					
member of B	I	II	III	IV	V	VI
1	0.73684	0.57895	0.57895	0.63158	0.52632	0.52632
2	0.73106	0.57895	0.57895	0.63158	0.63158	0.52632
3	0.73106	0.57895	0.73684	0.63158	0.63158	0.52632
4	0.73106	0.57895	0.73684	0.84211	0.68421	0.52632
average	**0.73684**	**0.57895**	**0.65790**	**0.68421**	**0.61842**	**0.52632**

Table 3.4: Probability of response B, given each of the members of category B for each of the six category structures according to equation 3.23 and the relative similarity measures given in table 3.3. The bottom row shows the average probability.

The reason why the order suggested by the confusion matrix is not precisely simulated by the configural-cue network model (see section 3.2.2.1.1), is due to the way in which learning takes place with this form of representation. While learning is dependent, to some extent, on these similarity measures, this is only the case because high intra-category similarity and low inter-category similarity corresponds to larger numbers of valid features and feature combinations.

As will be discussed in the next chapters, learning rates in the configural-cue model may not be perfectly described in terms of simple relationships between the quantity, frequency, and validity of its representations with respect to a particular task. While types III to V structures have more valid representations than the type II structure, the different logical status of stimuli within the type III to V structures means that there is a strongly interactive character to the relationships between them.

In relation to the identification and classification of continuous-dimension stimuli, the basic configural-cue model clearly has something of a problem. This has, to some extent, been addressed by a variant proposed by Shanks and Gluck (1994) known as the consequential region model. This model is based on theories summarised in Shepard (1987) regarding the mathematical modelling of generalisation.

According to Shepard's theory, a stimulus may be described as having what is described as a consequential region. This region may be represented in terms of a probability distribution with respect to its elicitation of a particular response. The approach is similar to the Luce model described above, and other theories relating classification and identification learning to the establishment of a distribution of response probabilities across the 'stimulus space' (e.g. Fried & Holyoak, 1984). Learning establishes the shape of this distribution and determines the degree to which the consequential regions of different stimuli overlap.

While the approach is most clearly represented by an exemplar-similarity approach, Shanks and Gluck (1994) suggested that one might achieve the equivalent by simply 'quantizing' stimulus dimensions and allowing multiple detectors per dimension. These detectors, with a finite range, may or may not have overlapping receptive fields and the receptive fields may be of different sizes relative to the region of interest for the experiment.

While this model may be fairly 'expensive' with respect to the number of detectors actually required, it was demonstrated by the authors to allow similar performance with respect to modelling identification and classification data to exemplar network models. This 'overlapping consequential region' approach has also been explored in relation to another, more recent model of categorisation given by Tenenbaum and Griffiths (in press, also Tenenbaum, 1996). Tenenbaum and Griffiths' model is principally a Bayesian model of category learning rather than a connectionist one, but it is fairly similar to the Shanks and Gluck (1994) consequential region model. It has, similarly, been applied successfully to the representation of a range of category leaning data involving continuous-dimension stimuli.

More research is indicated with respect to the application of this approach to modelling category learning, particularly in relation to situations in which selective

attention to dimensions is implicated (as will be discussed below). It does suggest that early reservations, regarding the ability of configural-cue networks to represent generalisation across continuous-dimension stimulus spaces, may be readily addressed by a reconceptualisation of the nature of the detector or stimulus representation. It is important to note, however, that for stimuli which are very similar to one another, such that one may not be able to reliably discriminate between them, the same problems apply to these approaches as to the exemplar network.

Another related problem would be that the number of consequential regions per dimension must be limited in some way in order to 'control' learning rates. The larger the number of simultaneously active detectors on a given trial, with a constant learning rate parameter, the faster learning will be. Using a Rescorla-Wagner learning rule will result in 'instant' asymptotic learning, if the learning rate parameter is greater than or equal to one over the number of simultaneously active representations. The resolution offered by this model may, therefore, be limited and also inflexible (without also assuming some change in the learning rate parameter).

3.2.3. Generalisation between stimuli with different numbers of features

Numerous observations in category learning research pose problems for connectionist models based on either exemplar or configural-cue representations of stimuli. The relationship between these problems and the mode of representation used in the model is, in many cases difficult to determine. As will be discussed in the next section, some of these problems appear to find solutions in the form of various algorithmic augmentations of basic models, incorporating processes which may be described as selective attention. The incorporation of these processes to particular models, however, does not necessarily enable a judgement to be made regarding which form of representation is most suitable.

3.2.3.1. Component and configural control over responding

Despite the success of exemplar based approaches in modelling category learning, a number of observations from research on associative learning appear to pose difficulties for this mode of representation. One of the main problems with the exemplar approach is its method of describing the generalisation of response strength between component and compound stimuli.

The exemplar approach appears to suggest that each stimulus, be it a compound of stimulus components, or simply those components presented in isolation, if it *is* presented and the response reinforced, should be represented by a separate exemplar. This is similar to the configural-cue approach, but the role which these exemplars play in control over responding is often modelled in different ways, depending on the data being simulated.

The observation of transfer to compound from component would appear to require that the component exemplar is present at the same time as the compound representation. The compound, having not being presented, will have no associative strength of its own and as such the transfer observation would have to be accounted for by the influence of the component exemplar.

The representation of transfer from compound to component, using the exemplar approach, is generally achieved in exemplar networks by assuming that the compound exemplar mediates responding, but that the similarity parameter for the absent component of the compound is, effectively, increased to one. In terms of the exponential similarity calculation given in equations 3.27 and 3.28, it requires an additional parameter to the distance measure for each feature (equation 3.26). Ths parameter multiplies the distance by zero if the feature is absent. While the similarity is still an exponentially decaying function of the distance between the exemplar and the stimulus, the distance for an absent feature is represented as being zero, such that exemplar activation is just a function of those features which are present.

What this effectively means, is that the activation of a compound detector given a particular stimulus is unity when the stimulus presented may be described as a subset, of any size, of the compound (e.g. Estes *et al.* 1989, Kruschke, 1992, and Nosofsky *et al.* 1992). The implication of this is that transfer to compound, following learning trials with a component, should be perfect. The other implication is that when training takes place on a single feature x and on a compound xy, both detectors will be active when x alone is presented.

This interpretation of the model was invoked by the above authors to enable the context model and the exemplar network model to try to account for a phenomenon observed in associative learning known as base-rate neglect. Base-rate effects are observed in category learning experiments when one category occurs more frequently than another.

In a classic experiment by Gluck and Bower (1988a) (replicated by Estes *et al.* 1989), for example, category A occurred on 0.25 of the trials and category B on 0.75. One of the features (of the four features used) occurred with a frequency of 0.6 if the category was A and 0.2 if the category was B. With respect to the feature itself the probability of category A equals the probability of category B given the presence of the feature. If shown the feature in isolation then the normative probability p(A| feature) = p(B | feature) =0.5. Participants typically, do not take into account the base-rates of the categories when assessing the probability of membership, and assign a much higher probability to p(A | feature) than p(B | feature).

A simple component cue network accounts for the effect in a straightforward way, due to the interactive behaviour of Rescorla and Wagner's (1972) learning rule. Because the cue above is the *best* available predictor of category B and the *worst* available predictor of category A, it tends to develop a higher associative weight with respect to B than it does to A.

Interestingly, neither the configural-cue model nor the exemplar network is particularly good at modelling this effect without substantial modifications (Nosofsky *et al.* 1992, Estes *et al.* 1989, Landenowski 1995). Base-rate effects will be discussed in more detail in the next section as models which appear to be able to represent them seem to require some form of selective attention.

Returning to the exemplar interpretation of transfer to and from components, the following situation appears to obtain. In order to explain transfer from component to compound, both the component and compound detectors must be active at the same time when the compound is being presented. In order to account for the generalisation of response strength from a compound to a component, however, the compound is described as being fully activated by any subset of its components. Recalling the observation from 3.1.3.1, that training on x to A then xy to A followed by subsequent presentation of y alone seems to indicate blocking of any learning about y and A, constraints are imposed on any exemplar based explanation of this process.

This would imply that during the training on x to A, an x exemplar is instantiated which then accumulates high levels of associative strength with respect to A. Presentation of xy paired with A should instantiate the xy exemplar whilst at the same time activating

the x exemplar to enable transfer. Because compound to component transfer is full in the exemplar model, the activation of x when xy is present must be sufficient to leave little or no error in the prediction of A given xy. If there were a 'gap' some associative strength would accrue to the xy exemplar. If this occurred then presentation of y in the following test phase would result in full transfer of response strength to y, which is a subset of xy.

This account of blocking results in serious problems for the model being able to account for the ability to learn a compound-component discrimination such, as x to A and xy to B as, once instantiated, both representations will be equally active on each trial. Creating an asymmetry with respect to the similarity of compound and component would appear to be justified here.

In order to allow the component-compound discrimination, the most practical step would probably be to allow the complete activation of the component in the presence of the compound, but to attenuate or eliminate the activation of the compound in the presence of the component alone. This, of course, would remove the ability of the model to account for compound to component generalisation. The most straightforward way of accommodating this may be to suggest that x and y exemplars are instantiated by the presence of an xy compound, and may acquire associative strength independently. The result is clearly a variant of the configural-cue model.

Yet another observation from associative learning, however, poses problems for the configural cue model (and consequently any variant of an exemplar model able to represent the various transfer, discrimination, and blocking results described above). Shanks, Charles, Darbi, and Azmi (1998) reported, with human participants, that a training phase on x to A and xy to B, followed by a training phase on y to A, did *not* interfere with the previously learnt x and xy discrimination when these were presented in a subsequent test phase.

This observation is problematic for configural-cue models, as these predict that a proportion of the association between xy and B will be a result of the accumulation of response strength for B by the y representation. The reversal of its association during the y to A training phase should, according to the model, have a serious affect on the response probability p(B| xy).

Shanks *et al.* (1998) interpreted this result as favouring an exemplar based approach, but *only* if that approach involved compound and component representations having activation functions which registered that x and xy were dissimilar. To offer an explanation, an exemplar model would have to assume that when xy was present, x and y component exemplars were much less active than when their component alone was present. When x or y was present, the xy exemplar would have similarly attenuated activation (*ibid.*).

3.2.3.1.1. Pearce's augmented configural-cue network

A model which offers this type of representation is a form of hybrid between exemplar and configural-cue theories and is best captured by that proposed by Pearce and Hall (1980) (also in Pearce 1987, 1994a, 1994b, Pearce & Redhead, 1993). This model was developed to account for the observation that when learning a stimulus-response pairing of the form x to A, xy to A and xyz to B, learning of the x to A pairing was more rapid than learning of the xy to A pairing.

The basic configural-cue network model predicts that the learning of xy to A should be more rapid because the x and the xy representation will be active during xy trials. For the x component alone, only one 'detector' is active and so the response strength should be less (Pearce, 1994a).

In order to account for this, Pearce's model proposed a form of activation for detectors based on the number of features each had in common with the presented stimulus. In the case of the above stimulus set, xy and xyz, for example, should be more similar to one another than x and xyz because they have more components in common. This should lead to greater generalisation of learning signals and consequently attenuate learning.

The activation function in Pearce's model provides yet another means of implementing the exponential decay of similarity with feature or Hamming distance. In the Pearce model (Pearce, 1987) the activation of a detector A, with A components, given the presentation of stimulus B with B components, or a(A | B) may be given by the following;

$$a(A|B)=\frac{A\cap B}{A}\times\frac{A\cap B}{B} \qquad (3.30).$$

Where A∩B is the number of features A and B have in common. An alternative formalisation of this similarity function was given in Pearce (1994b),

$$a(A \mid B) = \left(A \cap B \left(\frac{1}{\sqrt{A}} \times \frac{1}{\sqrt{B}} \right) \right)^2 \qquad (3.31).$$

If the detector is regarded as representing a vector in a feature space, with a co-ordinate of 1 for the presence of a feature and zero for its absence, the similarity function is basically the cosine of the angle between vector A and vector B squared (*ibid.*). The activation function is basically symmetrical, in that the activation of x in the presence of xy is the same as the activation of xy in the presence of x.

Learning in Pearce's model is effected by a variant of the Rescorla-Wagner learning rule. The exception is that, despite the total response strength being a sum of the activity of all detectors on a given stimulus presentation, associative strength only accrues to the 'focal' detector, i.e. the one that represents perfectly the current stimulus (*ibid.*).

The model can therefore predict the discrimination problem described above as well as compound to component and component to compound transfer, albeit at an attenuated rate compared to the basic configural cue model. Attenuation of this transfer in either direction has been a feature of its observation since first investigated by learning theorists (see, for example Hull, 1943, chapter XIII). Transfer to component is generally observed to be greater than half but less than full, with transfer to compound being less than double but greater than either component alone (*ibid.*). This attenuation is, significantly, not predictable by the basic configural-cue model.

The model also predicts that learning about a relevant redundant cue, as in the blocking paradigm, will be attenuated, although it does not appear to be able to represent complete blocking due to the symmetricality of the similarity function. Whatever does not generalise from x to xy will be learnt by xy and generalise, to some extent, from xy to y. This problem, as pointed out by Shanks *et al.* (1998) and Shanks, Darby, and Charles (1998), also appears to restrict the ability of this model to account for the resistance to interference effects described for humans in the previous section.

In addition, Pearce's model will be similarly unable to predict performance on the Shepard *et al.* (1961) category learning tasks. With this type of stimulus set, where all

stimuli have the same numbers of features, the model should behave in a similar manner to the exemplar network and the basic configural cue network.

These observations of Shanks *et al.* (1998) and Shanks, Darby, and Charles (1998) would appear to be quite damaging for all of the connectionist models presented above, particularly if they are to be used to predict other basic findings in human associative and category learning research. Whether they can be explained by modification of the models to incorporate selective attention learning processes or through use of alternative forms of representation seems uncertain. This resistance to interference appears to imply that there may be an extra 'dimension' to learning in the form of what these authors refer to as elemental versus configural processing.

It would appear that human performance on certain tasks indicates that sometimes people's learning may be best described in terms of elemental or component representations. For other tasks, such as in the resistance to interference tasks described here, people appear to use a strategy indicating that configural or 'whole stimulus' representations are being used. At the moment there do not appear to be any models capable of offering a principled account of what factors are likely to promote one type of learning over another (Shanks *et al.*, 1998, p.1377).

3.2.4. Stimulus and task dependent representations?

The foregoing summary of methods of representation employed in connectionist models of learning, indicates a wide array of techniques available to address various experimental findings. It may be argued that many of the models described above were not specifically developed to explain the observations of Shepard *et al.* (1961), with respect to the relative difficulty of their category learning tasks. Similarly, models such as the exemplar network model were not specifically developed to account for simple associative learning experiments where stimuli consist of collections of 'elements' and vary according to the presence or absence of these features (Kruschke, 1996a, p. 23).

Connectionist models may be said to begin with the hypothesis that 'events' may become associated with reinforced responses, according to some fairly simple rule describing changes in response probability as a function of the frequency of reinforced responses contiguous with that event. This 'rule' may be described in a number of ways. The methods, described in section 3.1, include simple response probabilities, or the

accumulation of some measure of response strength which may be readily converted to response probabilities for the purpose of making quantitative predictions.

One thing which appears to distinguish connectionist models from other mathematical models of learning, is their commitment to some form of theory regarding which specific characteristics of an event may get associated with a particular response. This commitment, as may be apparent, is by no means a straightforward undertaking. It requires one to propose what type of 'intra-systemic' event will accumulate and retain control over responding across learning trials and, for some experiments, across concurrent tasks. It is also an essential commitment in that the connectionist approach requires that each discriminable event, when a contiguous response is reinforced, be capable of exerting some form of independent control over responding.

The well-documented phenomenon of generalisation adds further complications to this commitment. It would appear to require that whichever way one decides to represent this intra-systemic event, its control over responding would appear to be a function of some measure of its similarity to other events which may or may not be explicitly represented in the model. The various models described above deal with generalisation in different ways. Sometimes this depends on the nature of the stimuli being represented; sometimes it is based on other theoretical considerations.

For simple stimulus events, such as those which may be described in terms of the presence or absence of certain features or elements, generalisation may be described in terms of some measure of what the stimuli have in common and/ or what is different about them. Set-theoretic models such as Amos Tversky's contrast model (Tversky, 1977, Gati & Tversky, 1982, Sattath & Tversky, 1987) represent a useful framework for analysing similarity in these terms. The basic expression of similarity in Tversky's model is given in terms of the similarity of stimulus a, with feature set A, to stimulus b, with feature set B, or S(a,b) as follows;

$$S(a,b) = \theta f(A \cap B) - \alpha f(A - B) - \beta f(B - A) \qquad (3.32).$$

Where S and f are interval scales and θ, α, and β are parameters greater than or equal to zero (Tversky, 1977). A-B consists of the number of features in a that are not in b, or the distinctive features of a.

Tversky suggest that this theorem does not define a single similarity scale, but rather a family of scales dependent on the values of θ, α, and β (*ibid.*). If, for example α and β are set to zero then similarity may be just be a function of the features that a has in common with b. Such a function is described by the configural-cue similarity function given in equation 3.29 and the similarity function used by Pearce in his augmented configural-cue model given in equation 3.31. If the similarity of stimulus, a, to a 'stored' stimulus representation for b is defined in terms of the distinct features, then similarity will be a function of the features that are in b but not in a, i.e. B-A. In this case a function representing this might be the distance metrics used in the exemplar similarity functions described above (Gati & Tverski, 1982).

Tversky's main point seems to be that the similarity function itself may depend on the stimuli being compared and on the type of decision which is dependent on this measure, such that a single function may be inappropriate for representing the entire spectrum of findings in generalisation studies. Whether this means that more than one model of stimulus representation is required, or whether the generalisation function may be, in some way, selected according to task constraints is uncertain.

One thing that seems quite clear is that the representational schemes proposed above, when implemented in terms of a connectionist network with basic learning schemes, are inadequate to explain numerous observations. While the Shepard *et al.* (1961) tasks provide one indication of this inadequacy, further constraints are provided by other learning tasks.

The Shepard *et al.* (1961) tasks are quite limited in terms of the constraints they impose on connectionist models of representation and learning. The stimuli used, for example, all have the same number of features or dimensions and the categories have the same frequencies. No transfer stimuli are presented such that a model capable of predicting the task difficulties does not necessarily have to offer any theories regarding how a participant may respond to new stimuli following learning of the tasks.

As discussed above, the representation of these stimuli for the purposes of deriving an identification-confusion matrix may be achieved using a number of schemes. Without further modification each predicts a similar erroneous ranking of task difficulty. Each of the forms of representation, however, may be differentiated according to the predictions

they make with respect to other tasks. This is frequently a function of the way in which each model represents generalisation.

As will be discussed in the next section, numerous experiments, including the Shepard *et al.* (1961) tasks, seem to indicate that generalisation is a flexible process. For example, generalisation does not operate equally with respect to all of the dimensions of a stimulus. Where the stimuli are not specifically dimensional and are based on features, then this process may be described in terms of a requirement to represent the differing 'salience' or diagnosticity of different features with respect to the extent to which they inform decisions made about them.

In a sense the problem is one of controlling generalisation according to the requirements of the task. This control appears to be something that is learnt. This requires that either a) some modification is made to the basic learning algorithms employed or b) 'secondary' learning processes are proposed which operate in parallel to basic learning algorithms and control dynamic parameters within a basic learning algorithm.

The control of generalisation is, obviously, a function of the way in which this generalisation is represented in the first place. As such the methods of control will depend on the method of stimulus representation employed.

Another feature, which is important from a modelling perspective, is that allowing a model to represent adaptive variability in its mode of generalisation provides it with new degrees of freedom with respect to its performance on all tasks. As with all cognitive modelling, the new capacities added to a model should not compromise its ability to perform as well as the un-augmented model, unless it can be demonstrated that the new capacity will not be manifested in the performance of the 'old' task.

3.3. Variability in the associability of stimuli

As discussed above, models of learning and categorisation are typically designed with a particular set of experimental data in mind. An analysis of the differences in the number and levels of independent variables between experimental designs must therefore inform the expectation that the performance of these models should generalise to other tasks.

Numerous factors have been identified which can have an effect on associative learning. The similarity of the stimuli requiring different responses to one another is one independent variable which may be reliably related to learning rate. Representing this effect in a connectionist model would appear to be dependent on the use of appropriate representations of the intra-systemic events resulting from the presentation of a stimulus.

The particular models of 'detectors' employed in connectionist models addressing generalisation phenomena are such that they are responsive either individually or *en masse* to the information which predicts generalisation. The identification and characterisation of a statistical relationship between independent and dependent variables is, in a sense, a basic model of that relationship (e.g. Estes, 1991). It may simply describe the shape of that relationship, as with the exponential growth function describing the relationship between frequency of reinforced responses to a particular stimulus, and the probability that the response will follow the stimulus on a given trial. The connectionist model attempts to suggest what kind of representations and information transmissions may be responsible for the observed curve.

The characterisation of connectionist architectures as communication systems will be explored in detail in the next chapter. It is important, however, to emphasise that the identification of a relationship between an independent and dependent variable may be interpreted as identifying a source of information for the mediating system. It is, to use biologist Gregory Bateson's definition of information, 'any difference that makes a difference' (Bateson, 1979, p.228). In order for this difference to make the kind of difference it is observed to make in experimental participants, one's model of the participant's processing must be capable of being affected in some way by the difference.

Sometimes this extra source of information may not have to be explicitly represented by a 'dedicated' process or variable in the model. For example, certain base-

rate effects, described briefly above, may simply manifest their effect via the behaviour of the same learning algorithms and representations used to model certain tasks where category frequency is equal.

One might regard the ability of a model to capture observations, such as that of base-rate neglect, as an 'emergent' property of the model designed specifically to model other effects. Unfortunately, as also mentioned above, the component cue models, for which base-rate neglect appears to be an emergent property, are incapable of representing learning of component-compound discriminations. For models which are capable of this, the kind of base-rate neglect observed in experimental participants is *not* an emergent property (Nosofsky *et al.* 1992, Landenowski, 1995).

As will be discussed below, the modelling of the effects of certain stimulus characteristics or differences in tasks, using a connectionist model, may require explicit representation, in some form, of the variable in question. One example of this may be the perceived intensity of a stimulus. Without some means of representing intensity differences it seems likely that it would be impossible for a model to capture the ability of experimental participants to be able to discriminate between stimuli with different intensities.

For others such as task complexity in the Shepard *et al.* (1961) category structures, the differences between category structures consist of the different relationships between stimuli within and between categories. While one could suggest some objective measure of the complexity of a category structure which correlates with subjective difficulty, it seems unlikely that this variable is going to be explicitly represented anywhere in the cognitive architecture in such a way as to 'control' learning rates.

Examples of operational definitions of complexity may be Shepard *et al*' s (1961) measure of the distribution of information across the stimulus dimensions, or Feldman's (2000) index of the Boolean complexity of a category structure. The former measure was discussed briefly in the previous chapter and will be described in more detail in the next chapter. Feldman's measure involves expressing the criterion for category membership within a structure in terms of a propositional concept. It is basically a formal version of Shepard *et al*'s (1961) assertion that difficulty was a function of the minimum length, in

number of clauses, of the rule required to describe how to assign any stimulus to a category.

This example highlights the differences between the connectionist approach and that used in the stochastic learning models described in section 3.1.2. If one can identify a 'quantity' using some analysis of the task which correlates with its subjective difficulty, one can propose a model of the type described in section 3.1.2 which is capable of representing the different learning curves. For example learning curves for each of the Shepard *et al.* (1961) tasks which reflect their subjective difficulty may be plotted using the following function where the change in the probability of responding 'A' given a member of category A, or 'a', as a stimulus is produced as follows;

$$\Delta p(A|a) = \theta(1 - p(A|a)) \qquad (3.33).$$

Where θ is either a greater-than-zero decreasing function of the Boolean complexity of the task (after Feldman, 2000), or a decreasing function of some measure of the distribution of information across the stimulus dimensions for the task.

This approach is not particularly helpful for producing connectionist models capable of capturing the subjective difficulty of a category learning task. In this case, the interesting questions are how something like task complexity may affect learning? What kind of generalisable representations and/ or learning rules are indicated by the fact that complexity *does* seem to have the effects it does on learning?

3.3.1. Factors affecting the associability of stimuli

Some of the factors which affect stimulus associability may be conceptualised as dimensions along which stimuli may, themselves, vary. It would seem that these differences might be most appropriately represented at the level of the stimulus representations. This would appear to be a problem of working out what kind of representations can facilitate the modelling of observed discriminatory performance across these dimensions. At the same time it may be the case that the *values* occupied by stimuli on certain dimensions may affect the rate at which they get associated with particular responses. Providing a model, particularly a connectionist model, capable of capturing both observations would appear to be a far from straightforward task.

Some factors relate to prior knowledge regarding the stimuli or previous contexts in which the stimulus has appeared. In this case the problem of modelling 'previous events'

can only really be approached, in a generalisable way, by looking at concurrent learning tasks in which the previous experience of a participant is, to some extent, 'known' by the experimenter.

Other factors may be dependent on the context in which a stimulus or stimulus component occurs. These may include, for example, the structure of the entire task or the relationship between the stimulus and other stimuli presented on the same or previous trials. In this case the focus of modelling may be on ways in which the stimuli are represented in a model, the way they all (including the context) are associated with particular responses, and the way in which these associations are maintained or not between trials or tasks.

Frequently in learning theory these various factors tend to be grouped under the label of 'salience'. Increasing salience implies increasing associability, with salience taken to mean whatever independent variable was increased, resulting in an increase in the rate at which stimulus was observed to get associated with a response.

Tversky (1977) suggested a division of a stimulus or feature's salience into two types of factor that he labelled 'intensive' and 'diagnostic'. Intensive factors may be described as those which pertain to what might be described as the perceptual aspects of the stimulus such as brightness, loudness or clarity. Diagnostic aspects pertain to what Tversky describes as the 'classificatory significance' of a stimulus or stimulus component (*ibid.* p. 342). The diagnosticity of a stimulus is equated with the prevalence (or significance) of classifications which are based on that stimulus or component (*ibid.*).

The representation of intensive factors is somewhat beyond the scope of this thesis but provides an interesting challenge for models of representation. The way in which they are represented appears, to some extent, to be a function of the experiment being modelled. For some experiments it appears to be appropriate to represent salience as something that has an effect on algorithms controlling stimulus associability. For others, such as the psychophysical studies described in section 3.1.2.2, intensive salience ought to be regarded as a dimension across which stimuli vary.

Diagnosticity, however, appears to be the focus of numerous models in both categorisation and associative learning theory. The following section will deal, in the main,

with issues regarding the contribution of the diagnostic aspects of salience to models of learning and categorisation.

3.3.2. Stimulus-specific learning rates

Despite the widespread influence of the Rescorla-Wagner rule in learning theory, there are a number of experimental observations for which, regardless of the representations used, it is unable to offer a simple account. Some of these observations, such as the Shepard *et al* (1961), data relate to the ability of models to predict the relative rates at which tasks will be learnt. Others relate to the ability of models to predict performance across concurrent tasks or situations.

Of this latter type a number of observations from learning theory would appear to indicate that learning may not only involve the development of some measure of associative strength between a stimulus and a response. In addition it would appear that experimental participants appear to develop some 'knowledge' regarding the diagnosticity of a stimulus or even stimulus dimension, which transfers between concurrent tasks. When a stimulus has been diagnostic for a particular task, it would appear that learning about it on a subsequent task is enhanced. Similarly, if a stimulus has previously been irrelevant with respect to predicting a classification, learning a new task where it *is* relevant seems to be attenuated.

3.3.2.1. The conditioned stimulus pre-exposure effect and learned irrelevance

One frequently observed manifestation of this transfer of knowledge is known as the conditioned stimulus (CS) pre-exposure effect or latent inhibition. It is frequently observed that unreinforced pre-exposure to a stimulus attenuates learning about that stimulus when conditions are changed, such that it becomes a reliable predictor of reinforcement. This attenuation is observed relative to new stimuli introduced at the time reinforcement is introduced. It is also noted that the pre-exposed stimulus does not become a conditioned inhibitor for the reinforced response in that its presence does not, in any way, attenuate learning for new stimuli. The basic Rescorla-Wagner rule is not capable of predicting this phenomena (see Mackintosh, 1975, and Miller *et al.* 1995 for a review of the problem). As the evidence suggests that the pre-exposed stimulus does not act as an inhibitor, the only 'place' for its associative strength to go is 'zero' which will put it on an equal footing with any new stimulus introduced.

A similar observation is noted when stimuli are pre-exposed in a manner that is uncorrelated with reinforcement. Prior to testing relative to new stimuli, the stimuli are presented in a context where reinforcement is occurring but is uncorrelated with any of the stimuli presented. Subsequent learning about these stimuli, relative to new ones, is similarly attenuated. This phenomenon is generally described as 'learned irrelevance'.

As Wagner and Rescorla (1972), and Mackintosh (1975) have pointed out, the only way in which the Rescorla-Wagner learning rule can be made to account for these observations is to propose stimulus specific learning rate parameters. It is suggested that these will decrease in the event of non-reinforced or uncorrelated pre-exposure to a level below that of a new stimulus such that, relative to these stimuli, learning is attenuated.

The CS pre-exposure effect and learned irrelevance are extremely significant with respect to the use of the Rescorla-Wagner learning rule. If, in order to offer some explanation of it one has to propose that the values of α_i for individual stimuli may vary as a function of experience, then there seems little reason to suggest that α parameters should not change during the course of any other kind of learning. The conditions that one suggests as being responsible for the alteration of these parameters during, in particular, uncorrelated pre-exposure are just as likely to obtain during learning of other kinds.

As pointed out by Mackintosh (1975) and Miller *et al.* (1995) this opens up the possibility that many of the observations predictable by the unique features of the basic Rescorla-Wagner framework may be predicted in terms of variation in learning rate parameters. In fact, any use of the Rescorla-Wagner rule in which α values for stimuli remain constant ought, in a way, to include an account of *why* they are constant under these particular conditions.

3.3.2.2. Mackintosh's theory of attention and associative learning

Neil Mackintosh's 1975 paper 'A Theory of Attention: Variations in the Associability of Stimuli with Reinforcement' summarises numerous observations of the role of different aspects of both intensive and diagnostic aspects of salience in associative learning. It also offered insights, which are still relevant (e.g. Kruschke, in press a), into how salience may be represented in models of learning.

Mackintosh equated the learning rate parameter with a process involving attention. Similar to the stimulus sampling approaches of Estes (e.g. Atkinson & Estes, 1963) and

Restle (1955), the parameter was related to some dynamic probability that a particular stimulus would be 'reinforced' on a particular trial. Having these parameters alter on each trial was meant to represent the fact that associative learning not only involves the accumulation of associative strength between a stimulus and a response, but also involves a process whereby the *relative* relevance of stimuli with respect to a task is learnt. The hypothesis is that participants learn to ignore irrelevant stimuli and, conversely, pay attention to the relevant ones. While this process is approximated, under many conditions by theories such as the Rescorla-Wagner rule in terms of differential rates of learning as a function of diagnosticity, the CS pre-exposure effect and learned irrelevance indicate that associative weights can not be the whole story.

Mackintosh's theory regarding dynamic, stimulus-specific learning rate parameters was to suggest that α_i would alter as a function of how well, relative to other stimuli present on a trial, stimulus i was capable of predicting the outcome. Formally, the theory suggests that where λ is the asymptotic maximum conditionable on a particular trial (or the 'target' value) and V_i is the current associative strength of i with respect to the response, then,

$$\Delta\alpha_i \text{ is positive if } |\lambda - V_i| < \left|\lambda - \sum_{j \neq i} V_j\right| \qquad (3.34).$$

The right-hand difference measure is between λ and the sum of all of the other stimuli, j, present on that trial. Also,

$$\Delta\alpha_i \text{ is negative if } |\lambda - V_i| \geq \left|\lambda - \sum_{j \neq i} V_j\right| \qquad (3.35).$$

The model assumes that $0 < \alpha < 1$ and also suggests that the size of the change in α will be proportional to the difference between the two differences (Mackintosh, 1975, p. 287-288).

The basic idea behind the model is that at the point of reinforcement (when λ is determined) the discrepancy between each stimulus' unique contribution to response strength and λ is compared to the discrepancy which would result if it were absent. If this discrepancy is greater than or equal to that which would occur in the absence of the stimulus then the α parameter for that stimulus will decrease. The rationale being that its

contribution is, at best, redundant. The learning rate parameter will increase if the discrepancy for a particular stimulus, i, is less than that which would occur in its absence.

Mackintosh expressed reservations about decreasing α when the difference in discrepancies was zero. Given that he felt that the magnitude of the change in the learning rate parameter should be proportional to the differences in discrepancies it may be 'preferable' if the change was zero when the difference in discrepancies was zero. The decrease, however, seemed necessary if the CS pre-exposure effect were to be explained using the model (*ibid.* p. 289).

As Mackintosh explains, the model predicts the CS pre-exposure effect by assuming that a stimulus, when presented in the absence of reinforcement enters into the comparison relationships shown in equations 3.34 and 3.35 with a set of background or contextual stimuli. Because it does not signal anything that the background stimuli do not already signal, its learning parameter will decrease. In this case if the background stimuli are conceptualised as being more numerous than one then the learning rate parameter for the stimulus in question will decrease because its contribution will be less than the sum of these stimuli. In fact, the learning rate parameters for all of the background stimuli will decrease for the same reason. Also, if at least some of the background stimuli are present before the stimulus for which the pre-exposure effect is being tested, then the learning rate parameter will decrease for this stimulus because its contribution is, again, less than rather than equal to the sum of the other stimuli.

The necessity for the decrease based just on redundancy of the stimulus (i.e. where its contribution is equal rather than less) appears to depend on whether one conceptualises the background as a 'unitary' compound or a set of stimuli. As will be discussed below, however, it may also become a factor when dealing with learning in compound stimuli.

3.3.2.2.1. Alternative interpretations of observations from associative learning

It is important to note that Mackintosh's model is *not* a connectionist model. It is basically 'a program for a theory' (*ibid.* p. 295) which proposes a set of rules which might apply to the representation of dynamic learning rate parameters.

Mackintosh actually suggested that the incorporation of stimulus-specific learning rate parameters might remove the need for the Rescorla-Wagner rule as a theory of learning. In his descriptions of learning, the form adopted represented something of a

return to the Hullian concept where the increment to associative strength was a function of an individual stimulus's discrepancy from the target value or λ on that trial, rather than that of the combined response strengths.

In the Mackintosh model, observations such as blocking may be accounted for by the reduction in the learning rate parameter for the redundant relevant cue introduced in the compound. Because its contribution will be less than that provided by the already 'trained' component it should only accumulate strength until such point as its learning rate parameter is reduced to zero. This should attenuate learning about the redundant stimulus, without recourse to a single increment signal determined by summed response strengths.

Another frequently observed effect in associative learning is known as *overshadowing*. When two stimuli are equally intense, but one is more reliably correlated with reinforcement than another, very little response strength tends to accumulate for the less diagnostic stimulus. The Rescorla-Wagner rule predicts this in that the superior reinforcement schedule of the 'overshadowing' stimuli means that it acquires strength much more quickly. This reduces the overall error signal available to affect the less valid stimulus. The Mackintosh solution is that the less valid stimulus, due to its inferior reinforcement schedule, will generally not predict anything that is not better predicted by the more valid stimulus. Its learning rate will subsequently decrease as its *relative* irrelevance is learnt.

As mentioned above, Mackintosh's theory is more of a framework of speculations which may be used to inform or guide more specific models. It does leave a number of questions open with regards to the relationship between the proposed attentional parameters and certain problems of learning and representation.

Mackintosh proposes, for example, that associability weights may also have a role in controlling the contribution of stimulus representations to the decision process. As will be discussed below, this suggestion has tended to be adopted by more recent models that incorporate selective attention processes. In this case the associability parameter may be better described as a secondary weight with its own learning process.

As will be discussed in the next section, certain observations suggest that these associability weights may, in some cases be dimensional rather than simply related to individual stimulus representations. Mackintosh suggested that such an interpretation may

be required to account for the results of experiments concerning dimensional relevance shifts. Numerous researchers (e.g. Kendler & Kendler, 1968, Zeaman & House, 1974, Kruschke, 1996b) have noted that if the response assignments of multidimensional stimuli are changed at some point during training, the new assignments are easier to learn if the change involves the same dimension being relevant than if the relevant dimension is changed.

The shortcomings of Mackintosh's approach are to some extent acknowledged in the conclusion to his article. In the context of this analysis they most obviously relate to its ability to handle configural learning and discriminations. The theory seems to relate best to an elemental approach to associative learning. Equations 3.34 and 3.35, above, indicate obvious problems with the learning of a compound association. Because each component of the compound would be making the same contribution to response strength, the implication is that each one's learning rate should fall.

This cannot really be mitigated by making the change to learning rates zero when the contribution of the stimulus is equal to that of other stimuli present on the trial. The problem would simply return if the compound consisted of three components. In addition, if the requirement for configural representations is acknowledged, any compound of two stimuli may be represented by three 'detectors'. Each cannot have its own weight without all of those weights falling towards zero during any learning where all were equally valid. This would be particularly problematic if the weights were also determining the contribution to response probabilities.

As mentioned above Mackintosh's theory was only intended as a framework. Its most significant contribution, as will be detailed below, is the idea that the utilisation of a particular stimulus or stimulus dimension is likely to be a function of its diagnosticity *relative* to other stimuli present on the trial. Results from concurrent learning tasks such as those involving learned irrelevance and relevance shifts seem to require the operation of a second learning process that is dependent on the relative validity of stimuli or stimulus components. Taking into account such processes also appears to be central to any attempt to model category learning experiments such as the Shepard *et al.* (1961) tasks.

The ways in which these processes may be implemented in a connectionist model are likely to be substantially dependent on the way in which stimuli are represented. As

will be discussed below, there appear to be a number of ways in which such 'selective attention' processes can be implemented for even a single representational scheme.

3.3.3. Dimensional attention: The generalized context model and ALCOVE

One of the most influential models of attentional processes in category learning extends from the application of the intuitions of Shepard *et al.* (1961), in relation to the task-dependent control of generalisation in exemplar representations. As mentioned in the previous chapter, Shepard *et al.* attempted to model the observed difficulty of their six category structures in terms of what was later described as the mapping hypothesis (Nosofsky, 1984) (see section 2.2.4.2 and 3.2.2.3.2).

The failure of the approach to predict the observed difficulties prompted Shepard *et al.* (1961) to propose that some process of abstraction or selective attention might be involved. This process, it was suggested, would mediate the involvement of the different stimulus dimensions such that generalisation would be just a function of the dimensions relevant to the particular task.

3.3.3.1. Robert Nosofsky's Generalized Context Model (GCM)

In examining the mapping hypothesis, Robert Nosofsky (1984) investigated Medin and Schaffer's (1978) hypothesis, that selective attention of some description may play a role in determining the similarity parameters of the context model (described in section 3.2.2.3.1). In the context model these parameters were fixed. This 'fixity' was justified in terms of the assumed equal intensive aspects to the salience of stimulus dimensions.

Nosofsky (1984, 1986) suggested that participants may distribute attention across different stimulus dimensions in such a way as to optimise performance, that is, maximise their percentage of correct responses on a task. To incorporate the selective attention process in a formal model, Nosofsky proposed an augmented version of the Minkowski r-metric formula (the general version of which is given in equation 3.26). Here the psychological distance between stimuli i and j, d_{ij}, is given by the following;

$$d_{ij} = c\left[\sum_{k=1}^{N} w_k \left|x_{ik} - x_{jk}\right|^r\right]^{1/r} \qquad (3.36).$$

In this case the dimensions are indexed by the subscript k so x_{ik} is the value of stimulus i on dimension k. The parameter c is a scaling parameter which is meant to represent some

measure of the overall discriminability in the psychological space, its value being greater than or equal to zero, r is the distance metric used, described in section 3.2.2.3.2. The 'attention weight' for a particular dimension k is represented by w_k. Nosofsky makes the assumption that $0 \leq w_k \leq 1$ and $\Sigma w_k = 1$ (Nosofsky, 1986). As described in section 3.2.2.3.2, the similarity is then the exponent raised to the negative power of the distance.

The attention weights for a particular dimension mediate the influence of 'distance' on that dimension on the overall similarity function. Figure 3.5 illustrates the effect of changing attention weights on the activation functions of detectors in a two-dimensional space. Panel A illustrates the distribution of similarities to three exemplars where the dimensional attention weights are equal (0.5) showing equal decay of similarity with distance along either dimension.

The lower panel of figure 3.5 shows the effect of differential attention with, in this case, enhanced attention to y. Nosofsky's attention optimisation hypothesis (Nosofsky, 1986) proposes that these weights will adjust to minimise average error on a particular task. If a dimension is irrelevant to a particular categorisation then the weight for that dimension will fall. This will reduce the effect of distance across the irrelevant dimension on detector activation.

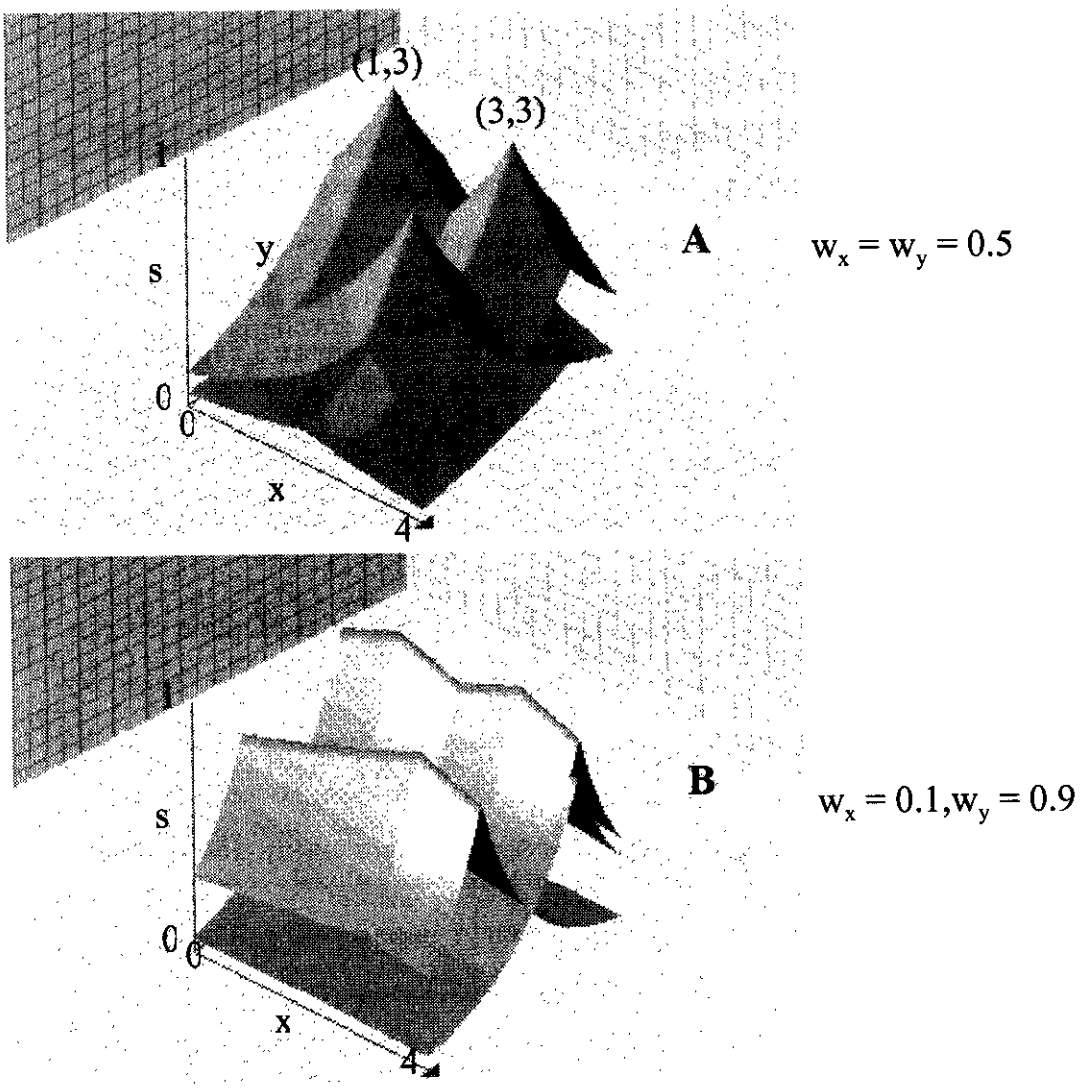

Figure 3.5: Distribution of similarity-based activation (s) for three exemplars located at (1,3), (3,1), and (3,3) in a two-dimensional space (x,y). Relative to equation 3.36 c = r = 1. Panel A shows an equal distribution of attention. Panel B shows the effect of extra attention to dimension y. Note decreased sensitivity to difference on the x dimension and steeper generalisation gradients along with respect to y.

With the Shepard *et al.* (1961) tasks, some dimensions are more relevant than others. In the type I task, for example, only one dimension is relevant. When the attention optimisation hypothesis is combined with the mapping hypothesis the suggestion is that attention to different dimensions will adjust to minimise the confusion errors with respect

to different values of the single relevant dimension. While limited capacity is not necessitated by the optimisation hypothesis, the limited capacity embodied by the normalisation of attention weights means that the GCM, with maximal attention to the valid dimension, will have zero attention weights for the two irrelevant dimensions.

In this case, each time an exemplar is presented to the network with this distribution of attention, all of those exemplars with the same value of the relevant dimension as the presented exemplar will activate maximally, and all those with the other value will be activated minimally. Effectively each exemplar is confused, completely, for every other exemplar with the same value on the single attended dimension.

Category structure	Optimal w values for dimension		
	q	r	s
I	1	0	0
II	0.5	0	0.5
III	0.35	0.3	0.35
IV	0.33	0.33	0.33
V	0.46	0.27	0.27
VI	0.33	0.33	0.33

Table 3.5: Optimal dimension weights for Shepard *et al.* (1961) category structures shown in figure 2.1 given c value of 5. Adapted from Nosofsky (1984).

Using a parameter optimisation procedure (see Nosofsky, 1984), Nosofsky calculated the optimum attention weights for each of the Shepard *et al.* (1961) tasks given a value for c of 5 (for equation 3.36). With reference to figure 2.1 the optimum weights for each of the dimensions q, r, and s are given in table 3.5.

These weights may be applied to the exemplar activation equations given in equations 3.36 and 3.28. Using these, one may construct a confusion matrix which may be collapsed, using the mapping hypothesis, to give the ratio for calculating choice probabilities based on the sum of similarities to members of one category over the sum of similarities to all exemplars (given for the context model in equation 3.23). If the attention

weights are different then the identification confusion matrix will be different. Note that the identification confusion matrices will be the same for types IV and VI where the weights are equal.

As with the context model and the configural-cue model, these matrices may be used to determine the average probability of confusing a given exemplar for a member of its own category. These are given in table 3.6 and, as can be seen, they now predict the correct ordering of subjective difficulty for each task.

	category structure					
member of B	I	II	III	IV	V	VI
1	0.99331	0.85979	0.72354	0.74996	0.60683	0.65879
2	0.99331	0.85979	0.72354	0.74996	0.87422	0.65879
3	0.99331	0.85979	0.85195	0.74996	0.77518	0.65879
4	0.99331	0.85979	0.85195	0.9323	0.77518	0.65879
average	**0.99331**	**0.85979**	**0.78775**	**0.79555**	**0.75785**	**0.65879**

Table 3.6: The predictions of the GCM in terms of probability of confusing a member of category B for a member of category B, given optimal attention weights shown in table 3.5 and a c value of 5. Bottom row shows average probabilities of confusing an exemplar for a member for its own category.

As mentioned above, the attention weights in the GCM are determined by a process of constrained parameter search. In the case of the Shepard *et al.* (1961) tasks the constraint involved is one that minimises predicted error by the model. This involves finding attention parameters that maximise intra-category similarity and inter-category differences. The goal being to minimise the predicted probability that a stimulus will be confused (according to the mapping hypothesis) for a member of the category it does not belong to.

Note that this parameter optimisation is for the basic context model which, as discussed above, suffers from the overlap problem. This may enable the subjective

difficulty to be predicted but it does not enable the generally perfect asymptotic performance of participants on the tasks to be predicted using the context model's learning rules. Except for the type I task, perfect or near perfect performance cannot be predicted by the GCM. As discussed above, a solution to the overlap problem for the context model may be developed by 'attaching' associative weights to the exemplars and allowing them to learn according to some variant of the Rescorla-Wagner learning rule.

The idea that attention weights will develop according to this constraint is the attention optimisation hypothesis. As discussed in section 3.1.3.3.1, learning rules such as the Rescorla-Wagner and Widrow-Hoff rule implement a similar constraint. Associative weights, when represented as a position in a 'weight space', generally head towards locations which minimise the mean squared error of the network (i.e. its discrepancy from the target values of its various outputs). The implication is that attention weights are learnt according to some similar constraint, at the same time as association weights between exemplar-based detectors and responses are developing.

3.3.3.2. John Kruschke's ALCOVE

One effective solution to the problem of modelling the 'learning' of attention weights within an exemplar based network was presented by John Kruschke in the form of the Attention Learning COVEring map or ALCOVE (Kruschke, 1992). This model proposes that attention weights are learnt at the same time as the association weights, developing according to a process of 'back-propagation' of error.

Standard back-propagation of error (Rummelhart, Hinton, & Williams, 1986) was developed as a means of overcoming the limitations of a standard component cue network or single layer perceptron with respect to linearly inseparable problems. The basic architecture proposes a layer of 'hidden units' between elemental input nodes and output nodes. The hidden layer nodes (which receive weighted input from all input nodes and are fully connected to the output nodes) develop what may be described as internal representations of the various features and feature correlations required to minimise mean squared error for the task.

Standard back-propagation will not be discussed in detail here. Despite the enormous popularity of networks based on this technique in certain areas of cognitive modelling, it is generally incapable of representing much of the data which emerges from

the study of categorisation (see Gluck, 1991, Kruschke, 1992, and Kruschke, 1993 for examples). The basic idea behind back-propagation, however, is that the layer of input units develop connection weights to those hidden units which are active at the same time as them, according to how well the hidden units predict the 'reward' signal at the output layer.

ALCOVE uses a variant of the basic back-propagation scheme to develop attention weights according to how well variation in their dimensional values correlates with variation in the category label. The basic architecture of ALCOVE is shown in figure 3.6, as 'set up' for the three-dimensional category structures of Shepard *et al.* (1961).

Figure 3.6 shows three dimensional attention weights. These weights modulate any distance measure across the dimension to which they are attached. The distance measure determining the activation of the exemplar detector in ALCOVE is the same as that given in equation 3.36 for the GCM. The activation, a, of any exemplar detector, i, given presentation of a stimulus, j, is given by the following function;

$$a_i = e^{-c\left[\left(\sum_x \alpha_x |i_x - j_x|^r\right)^{\frac{1}{r}}\right]} \qquad (3.37).$$

Here i_x is the value on dimension x detected by the exemplar detector, and j_x is the value on that dimension which pertains to the input stimulus j.

Activation from the exemplar detectors is passed via weights between each detector, i, and each alternative in the decision process, *l*. The total delivered response strength, to a label, or its activation, a_l is evaluated for each stimulus presentation according to the following function;

$$a_l = \sum_i a_i w_{il} \qquad (3.38).$$

Activations of the label nodes may be converted into response probabilities in one of two ways. In Kruschke's original presentation of the model (Kruschke, 1992), the response probabilities were determined using a logistic based function or ratio of the exponentiated activation of one node divided by the sum of exponentiated activations (as in equation 3.13). For a later application, which involved producing a quantitative fit to the replication of the Shepard *et al.* (1961) experiments, the response probabilities were the simple ratio of the activation of one label node over the sum of activations. Note this

requires that negative activations are truncated at zero, and a background noise or bias parameter be involved as described in section 3.1.2.1 (Nosofsky *et al*, 1994).

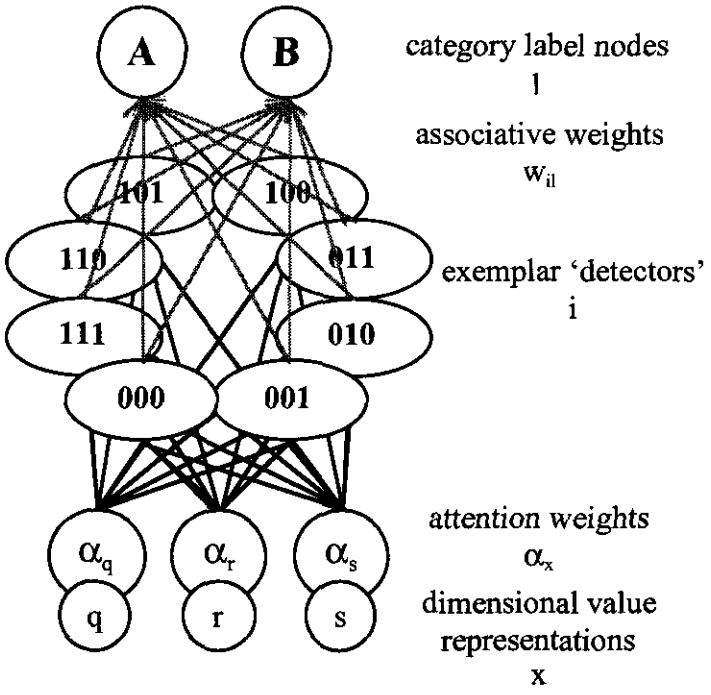

Figure 3.6: Basic architecture of ALCOVE as set up for the Shepard *et al.* (1961) tasks with three-dimensional input, one exemplar detector, i, for each of the eight patterns used, and two category label nodes, *l*. Notation used follows that in the text.

3.3.3.2.1. Associative learning in ALCOVE

For the changes in association weights between exemplar detectors and category label nodes, Kruschke employs a variant of the Rescorla-Wagner rule. The variation concerns the use of what Kruschke calls 'humble teachers' (Kruschke, 1992, p.24 & 39). The assumption behind humble teachers is that participants only receive nominal feedback regarding category membership and do not receive any information about degree of membership. To implement these intuitions Kruschke truncates the effect of the teacher signal to zero if the activation of the category label node exceeds one, where the stimulus was a member of that category. If the category label node's activation is less than minus one (or zero in the case of Nosofsky *et al.*'s (1994) 'background noise' version) when the

stimulus was not in that label's category, the teacher signal is also truncated at zero. Formally, the teacher signal for a label node *l*, or t_l given presentation of stimulus j, is expressed as follows (from Kruschke, 1992, where response probabilities are calculated using a logistic);

$$t_l = \begin{Bmatrix} \max(+1, a_l) \text{ if } j \in l \\ \min(-1, a_l) \text{ if } j \notin l \end{Bmatrix} \qquad (3.39).$$

This is incorporated into the following learning rule which determines the increment to the weight between exemplar detector i and label node *l*;

$$\Delta w_{il} = \lambda_w (t_l - a_l) a_i \qquad (3.40),$$

where λ_w is a learning rate parameter which applies to the association weights in the model.

As Kruschke (1992) points out, the use of humble teachers does not make that much difference to the predictions made by ALCOVE. Nosofsky *et al.* (1994) suggest that when examined quantitatively (with free parameters optimised to minimise discrepancy between the model's predictions and experimental data) the humble teachers do result in slightly better fits to data.

It ought to be noted, however, that the category label activation (or any measure of theoretically unbounded summed response strength) is only related probabilistically to the occurrence of a particular response. It may be the case, therefore, that the magnitude of any teacher signal may be better related to the probability of a particular response occurring than to the underlying response strength.

3.3.3.2.2. Attention learning in ALCOVE

The changes which accrue to attention weights are determined by what may be regarded as signals passed back via the exemplar detectors, to the dimensions on which they are dependent. The rule given by Kruschke (1992, p. 24) for changing the attention weight for a dimension x given presentation of a stimulus j is as follows,

$$\Delta \alpha_x = -\lambda_\alpha \sum_i \left[\sum_l (t_l - a_l) w_{il} \right] a_i c |i_x - j_x| \qquad (3.41),$$

where λ_α is positive learning rate parameter for the attention weights. If the attention weight drops below zero as a result of this change its value is clipped at zero. In the first

published report on ALCOVE (*ibid.*) attention weights were otherwise unbounded. The model was tested with respect to its ability to generate quantitative fits to the Shepard *et al.* (1961) category learning tasks by Nosofsky *et al.* (1994). In this model the incremented attention weights were normalised prior to the beginning of the next trial such that they summed to unity, as in the GCM (Nosofsky, 2000, personal communication). As will be discussed below, this normalisation is important for ALCOVE to be able to make the predictions it does regarding the Shepard *et al.* (1961) category structures.

The way attention learning operates in ALCOVE is fairly subtle. ALCOVE effectively tests each dimension for relevance on each trial where a teacher signal is present. The last component of equation 3.41 means that the attention weights only receive increment signals via those exemplar detectors activated by *generalisation*. For the exemplar representing the actual stimulus this last difference measure will be zero for all dimensions.

Objects activated due to generalisation receive error signals which have a size proportional to their activation and a direction which is a function of their output weight. These error signals are reversed in sign due to the $-\lambda_\alpha$ at the start of the learning rule. If the exemplar is predicting the *right* label it gets a negative error signal and if it is predicting the *wrong* label it gets a positive signal. This signal is passed back through the dimensional detectors where it is multiplied by the *distance* between the detector and the instantiated input. This is the distance multiplied by the total (not weighted according to current attention weight) specificity (the c parameter).

The net effect is that the 'correct' object representation does not affect attention weights as all of its distances are zero. Those activated by generalisation, however, do. If a detector is activated by generalisation across a particular dimension and is predicting the same label as that which has actually occurred, it passes a negative signal to the dimension weight. This 'works' because it is an indication (under the constraints of the particular experiment) that variation in the input on that particular dimension is not related to variation in the category label. If it was, then a different dimensional value should result in a different label.

The dimension weight gets a large negative signal if the detectors active as a result of generalisation across it are predicting the same label as that which has occurred, and a

large positive signal if they are predicting a different label. Positive signals increase the proportion of overall specificity 'used' which means that the size of subsequent signals to this dimension will fall, as there will be less generalisation of activation across the dimension. Negative signals mean that the dimension has a lower weight which means that activation will generalise even more across the dimension.

When applied to the Shepard *et al.* (1961) category structures the version of ALCOVE (Kruschke, 1992) with unbounded attention strength and logistic choice function readily predicted the observed ordering of difficulty. There were some problems with the 'shape' of learning curves produced by this model when compared with the curves produced from human data by Nosofsky *et al.* (1994) (shown in Figure 2.3). These 'fits' were improved by Nosofsky *et al.* (1994) and much of this improvement may be attributed to the use of a limited capacity assumption with respect to the dimensional attention weights.

Examination of Kruschke's original simulation results (Kruschke, 1992, p. 28) in relation to the later data from the Nosofsky *et al* (1994) replication of the Shepard *et al.* (1961) tasks indicates two potential difficulties. The first is the late manifestation of superiority for the type II task over the types III to V; the second is a somewhat close proximity of the type VI learning curve to the curves for tasks III to V.

These two problems are related and, as mentioned above, are both ameliorated by incorporating the assumption of limited capacity by using normalised attention weights. Beginning as it does with a number of detectors equal to the number of stimuli (each with their own associative weights) ALCOVE enhances the learning rates for tasks where one or more dimensions are irrelevant to the task via a form of recruitment. As the attention weight for an irrelevant dimension drops towards zero its ability to control detector activation decreases. For the type I task, with two irrelevant dimensions, this results in four detectors becoming active for each stimulus presentation (i.e. the four nodes with the same value as the stimulus on the relevant dimension). The rate at which the network learns the task thus approaches four times that of the initial network where only one exemplar was maximally active per trial. With the type II, as the attention weight for the irrelevant dimension drops towards zero, the number of detectors active on each trial approaches two.

The faster the attention weights alter, the faster this 'recruitment' advantage may be manifested.

When all dimensions are relevant to the task, the initial relative difficulty of the tasks will be a function of the generalisation relationships described in section 3.2.2.3.4. Using unbounded attention weights actually favours attention learning in the type VI structure. Examination of the category structures in figure 2.1 reveals that all of the nearest neighbours to detectors in the type VI structure are in the opposite category. As far as the simple stimulus generalisation model is concerned, this accounts for the advantage of the types III to V structures over the type VI. All generalisation for the type VI structure reduces the probability of selecting the correct category, whereas for the types III to V generalisation will sometimes result in confusion for a member of the same category.

With unbounded attention weights the generalisation problem for the type VI structure may disappear, particularly as learning progresses, in a way which is dependent on the size of the attention learning rate parameter. Equation 3.41 means that when detectors from the opposite category are activated by generalisation across a dimension, the attention weight for that dimension will increase. Kruschke (1992) reported that the attention weights for the type VI generally increased across learning whereas those for the type III to V remained close to the initialised values of 1/3. For these structures sometimes a dimension is relevant and sometimes it isn't.

Normalising the attention weights enables the attention learning rate parameter to be increased, thus enhancing the rate at which types I and II, with their irrelevant dimensions, are learnt but not affecting the rate at which the type VI is learnt relative to the types III to V. Because these four tasks require all dimensions, their relationships to one another will be largely similar to that displayed by a model with no attention weights.

Attention learning in the exemplar network model, it must be pointed out, does not necessarily enhance the ability of exemplar-based networks to simulate human experimental performance. Nosofsky and Palmeri (1996) applied the exemplar network model to a task involving the classification of *integral* dimension stimuli. As discussed in section 3.2.2.3.2, integral dimensions are those which appear to combine into relatively unanalysable wholes when present in the same stimulus. For these stimuli, experimental data is best simulated using exemplar-based models when the r metric in the distance

function (see equation 3.27) is set to a value of 2, facilitating an Euclidean distance measure.

Nosofsky and Palmeri (1996) generated experimental stimuli consisting of different patches of colour varying on the three dimensions of hue, brightness, and saturation. They then tested participants on learning tasks involving the same abstract category structures as those used in Shepard *et al.* (1961) to examine the order of difficulty displayed for these structures using integral dimension stimuli. In this case participants generally learnt the types III, IV and V structures with less errors than they did the type II structure, the order being I, IV, III, V, II and VI.

As discussed in previous sections this is the order predicted by the exemplar model when no processes of dimensional attention are operating, as given in table 3.3. Nosofsky and Palmeri (1996) subsequently attempted to fit ALCOVE to the data and discovered that the best fitting parameters, somewhat unsurprisingly, involved the attention weight learning parameter being set to zero. They regarded this as an indication that selective attention processes may be considerably attenuated when the stimuli concerned vary on integral dimensions rather than separable ones.

3.3.4. Modular approaches to attention

The approach to attention learning in the GCM and ALCOVE is to locate attention weights on the dimensions of the stimuli presented. This is one option with respect to the way in which the observed effect of attention may be mediated. Unfortunately, ALCOVE suffers from the same problems as the simple exemplar network when it comes to representing learning and transfer involving stimuli with different numbers of components.

Despite its effectiveness with respect to tasks such as the Shepard *et al* (1961) category structures and the dependence of this effectiveness on its selective attention processes, it is difficult to see how ALCOVE could address observations such as that of learned irrelevance, compound-component discriminations, base-rate effects, and even blocking of conditioning. The addition of ALCOVE's attention learning mechanism to the basic exemplar network does not seem to assist the model in relation to making plausible predictions about some fairly basic observations of associative learning.

A second family of models which formalise intuitions regarding processes of selective attention make use of a *modular* organisation. Unlike models such as ALCOVE,

attention or modular weights are, generally, proposed to lie between stimulus representations, or detectors, and the decision process.

These models were not motivated by the shortcomings of models such as ALCOVE with respect to representing things like learned irrelevance, blocking, and compound-component relationships. In some cases these models have their origin in more 'engineering' or abstract theoretical considerations. In these cases they have been imported to the study of learning on account of their ability to implement forms of attention using representations other than the exemplar-based detector. In other cases the models have emerged to address shortcomings of models such as ALCOVE in accounting for data from other category learning experiments.

3.3.4.1. Dynamic Learning Rate (DLR) models

One approach to implementing something like selective attention in a connectionist network is to propose, as Mackintosh (1975) did, dynamic learning rate parameters for each stimulus component or representation. A method of implementing these parameters in connectionist networks was proposed by Richard Sutton (Sutton, 1992) and implemented in the form of a psychological model first by Gluck, Glauthier and Sutton (1992). The algorithm, known as Incremental Delta-Bar-Delta (IBDB), was proposed by Sutton as a means of accelerating learning in networks using the Least-Mean-Square learning system by altering the learning rate parameter for a connection as a function of its history of changes.

The logic behind the algorithm was that if a series of weight changes for a particular connection are all in the same direction, then the weight could probably reach its asymptotic value quicker if the learning rate parameter was larger. If, on the other hand, changes are in alternating directions then the weight is probably overshooting its asymptotic value and as such a smaller learning rate may be justified (Gluck *et al.* 1992).

This is applicable to a learning task where some stimuli are more relevant to predicting the outcome than others. Stimuli for which successive teacher signals associated with the stimulus's activation are in the same direction are likely to be well correlated with the outcome. Stimuli for which error signals switch direction may be irrelevant to the prediction of the outcome.

The model represents the history of associative weight changes using a memory variable. This controls alteration of the learning rate parameter according to whether the current teacher signal is in the same direction as those which have recently occurred when the representation has been active. The model uses the same method for calculating the teacher signal as the Rescorla-Wagner rule.

Gluck *et al.* (1992) applied the algorithm to a configural-cue network design in order to represent category learning data, and found that it was better able to represent human data than the standard configural-cue model of Gluck and Bower (1988b) (see Gluck *et al.* 1992 for details).

One notable problem with the standard configural-cue model is that it tends to distribute response strength across a variety of nodes in a way that does not perfectly represent their individual diagnosticity with respect to the task. This is both a strength and a weakness of the Rescorla-Wagner learning rule.

It is a strength in that it is this which enables the basic configural cue network to be able to learn a compound-component discrimination task. Significantly the DLR model is likely to find this task fairly difficult, as the components will be receiving positive teacher signals when they are present alone and negative ones when presented as part of a compound. Their learning rates are thus likely to decrease due to their lack of correlation with the teacher signals for each outcome.

It appears to be a weakness in that it results in the model allocating too much weight to partially valid nodes, particularly prior to asymptotic levels of learning. This would appear to be the weakness located by Gluck *et al.* (1992) in their investigations, and also a weakness of the model with respect to the Shepard *et al.* (1961) tasks (Nosofsky *et al.*,1994).

3.3.4.1.1. Application of the DLR approach to the Shepard et al. (1961) tasks

As discussed in previous sections, the standard configural-cue model appears to perform badly with respect to predicting the subjective difficulty of the six category structures. One reason for this is that it appears to assign too much weight to cues and cue configurations which may be regarded as only partly diagnostic. As Nosofsky *et al.* (1994) pointed out, this was only a problem with respect to the partially valid one-dimensional cues for the types III to V structures. Most of the problem seemed to originate from the

fully valid two dimensional cues which, when totalled across the two or three 'spaces' in which they occurred, were more numerous than the number of valid cue configurations present for the type II task.

Reasoning that this was unlikely to be remedied by the DLR model with a separate learning rate for each cue and cue configuration, Nosofsky *et al.* (1994) tested a variant of the model which they called the Dimensionalized Adaptive Learning Rate Model (DALR) (*ibid.*). The difference between this model and the model described above is that while each connection has its own history parameter in the DALR, learning rates are shared by 'spaces' such that there are just seven adaptive learning rates (q, r, s, qr, qs, rs, and qrs) for each output node. These rates are described as linked for a space in that each update is dependent on the particular node which is active. The update, however, is to the shared learning rate parameter.

The best fits of this model, while producing a qualitative fit to the total error data, were fairly poor. The superiority of the type II task over tasks III and IV was only marginal and also emerged only after about three training blocks. Nosofsky *et al.* (1994) noted that the DALR still seemed to give too much weight to the doublet cues in the type III and IV structures. The reason for this may be predicted to some extent from the operation of the IBDB algorithm when coupled with the 'dimensionalisation'. In the case of the spaces where the fully valid doublet cues occur, each will have two cues which are perfectly valid and two which will not be valid at all. On half of the trials, therefore, the learning rate for the space will be reliably increasing whilst on the other half it has a 50% chance of increasing and a 50% chance of decreasing. The net result is likely to be an overall increase. The fact that the rate of increase will be less than for a fully valid space will be at least partially offset by the larger number of cues involved.

3.3.4.2. Mixture of experts (ME) models

Mixture of experts architectures comprise a 'family' of models which propose task decomposition and processing by sets of 'local' experts or modules (Jacobs, 1997). Modularity, to varying extents, appears to be a characteristic of nervous systems where certain areas appear to be distinct from others in terms of the sources of their inputs and destinations of outputs. This has motivated the specific designs of a number of so-called 'neural' models such as Adaptive Resonance Theory or ART (see, for example, Grossberg,

1987, Carpenter & Grossberg, 1993, for reviews) and the Categorizing and Learning Module or CALM (Murre, Phaf, & Wolters, 1992). In addition it has been demonstrated using connectionist networks, that dividing the architecture into modules with some degree of functional specialisation may enhance the ability and efficiency of those networks to learn certain tasks (e.g. Rueckl, Cave, & Kosslyn, 1989, Jacobs, Jordan, & Barto, 1991, Jacobs, Jordan, Nowland, & Hinton, 1991).

Despite the success of approaches such as ART in being able to represent observed cognitive capacities, they do not seem particularly suited to modelling the fine details of category learning data. The mode in which stimuli are represented in Adaptive Resonance Theory, for example, which is likely to involve unique representations for each stimulus with little or no generalisation between them, would appear to preclude the possibility of them being able to model the Shepard *et al* (1961) data. As such this section will concentrate on some of the approaches which have been specifically applied to category learning data.

The basic design of such models may be most readily related to that proposed by Jacobs, Jordan, Nowland, and Hinton (1991) (see Jacobs, 1997, for a review of various contributions to the model's design). In this scheme a set of modules containing detectors for different aspects of the input receive teacher signals *and* contribute to the decision process via a set of nodes described by Jacobs, Jordan, Nowland, and Hinton (1991) as a 'gating network'. This distinguishes the approach from that used in the DALR, as for the DALR the parameters between the modules and the decision process only mediate learning rates. This captures the intuition of Mackintosh (1975), described above, that attention parameters of some description may have to control the effect of connections on performance as well as the learning rates of those connections.

The gating network effectively selects whichever module is predicting the outcome most successfully as the principal contributor to the decision process. In the Jacobs, Jordan, Nowland, and Hinton (1991) model this process was stochastic, with the gating nodes' activation to each module determining the probability that that model would be the one which 'made the decision' on behalf of the system. Some variants of the model, e.g. Jacobs, Jordan, and Barto (1991), alter these probabilities based on whether the performance of the system is significantly improved by previous updates. Others, such as

an implementation developed by Erickson and Kruschke (1998), use adaptive weights which are modified by gradient descent on error between the nodes in the modules and the gating node. The latter variant is somewhat different from the Jacobs, Jordan, Nowland, and Hinton (1991) model in that the gating nodes' activation is deterministic rather than stochastic. In both cases the gating node activation is normalised in some way such that enhanced contribution from one module is at the expense of contribution from others.

The gating nodes also gate teacher signals to the nodes in the module they receive input from. The approach is also reminiscent of Mackintosh's approach, in that learning within modules is local to that module (Kruschke, in press a). For the stochastic implementation, in a situation where only one node per module is active at any one time, this is equivalent to learning based on the summed activation at the decision as the decision is only based on the activation of one module. For implementations, such as Erickson and Kruschke's (1998) model, the learning is actually local (although still gated by the normalised gating node activities).

There are a large variety of models based on this type of architecture, all of which employ slightly different learning, gating, and representational assumptions. The model of Erickson and Kruschke (1998) known as ATRIUM (Attention To Rules and Instances in a Unified Model), for example, used a full implementation of ALCOVE, including dimensional attention weights, for one module and a set of 'rule-representing' nodes as another module.

The model was designed to account for the learning of exceptions in a mostly one-dimensional rule-based category structure. It has been noted (e.g. Nosofsky, Palmeri, & McKinley, 1994, Palmeri & Nosofsky 1995) that models such as the context model or ALCOVE fail to account for the ease with which people learn structures which are generally defined by simple rules but include a few exceptions. ALCOVE and the context model would require attention to all dimensions for the model to be able to represent learning of the exceptions. This distribution of attention would significantly attenuate learning, relative to that which seems to occur, of the more frequent stimuli to which the rule applies.

Nosofsky, Palmeri, and McKinley, (1994) and Palmeri and Nosofsky (1995) proposed a form of ME architecture known as RULEX (rule-plus-exception model) which

involves mostly rule-based learning. This is supplemented by the occasional exemplar representation that gets 'instantiated' when the outcome of a particular trial is the opposite of that predicted by the rule. The model has a stochastic gating mechanism that decides whether the network is going to make the decision based on the rule or on the presence of an exception and, eventually, learns to pay attention to its exception representations at the appropriate times.

ATRIUM (Erickson & Kruschke, 1998) makes use of a module containing what are, effectively, component cue representations which are supposed to represent simple rules, and a high dimensionality (usually with as many dimensions as the stimuli used) module to represent exceptions. Which module the decision process 'pays attention' to when a particular stimulus is presented, depends on whether that stimulus has previously been an exception to the rule or not.

When an exception is presented,the model generally allocates most of the resultant error signal to the high dimensionality module, thus leaving any component cue rule representation largely intact. As such, large associative weights only acumulate to the sources in the high-dimensionality module which are representing exceptions. The decision process is thus able to 'prioritise' input from each of its modules. Because of the way the error signal is allocated to modules according to whether they conform to the rule or are exceptions, the model 'knows' that if it is receiving a substantial signal from the high-dimensionality module, then the stimulus is probably an exception.

It is not clear how models such as RULEX and ATRIUM would be used to model learning of the Shepard *et al* (1961) tasks. ATRIUM uses a full version of ALCOVE, complete with dimensional attention learning, as its mechanism for learning exceptions (Erickson & Kruschke, 1998). As such, it is difficult to see whether it makes any falsifiable claims about how catgeory structures with and without rule-plus-exception aspects would be learnt.

Generally the ATRIUM model is fit to experimental data using a parameter optimisation procedure. One of the estimated parameters used in the model is a 'gate bias' which decides, initially, how much 'attention' should be paid to the exemplar module relative to the rule, or component cue module. Using the process it is easy to see that the difficulty of the Shepard *et al.* (1961) tasks could be well represented by simply setting this

gate bias to a sufficiently high value to guarantee that ALCOVE does all of the work (Erickson & Kruschke, 1998, p.119).

3.3.5. Rapid attention shift models

A final class of models to be discussed here incorporate algorithms to generate rapid shifts in attention weights *after* feedback has been presented and before the associative weights are adjusted. This type of model was first developed by John Kruschke (Kruschke, 1996a) in order to model certain base-rate effects. Since then a number of models have been proposed by Kruschke and his colleagues incorporating the idea of rapid attention shifts. This section will first detail the base-rate effects that motivated the development of rapid attention shift models. Following this the variety of models incorporating rapid attention shifts will be discussed.

3.3.5.3. The inverse base-rate effect and base-rate neglect

Kruschke developed his model, called ADIT for Attention to Distinctive Input to model an effect known as the *inverse base-rate effect*. The effect, noted by Medin and Edelson (1988) involves participants tending to make a rare category assignment when presented with a pair of cues, one from each of the rare and common categories.

The replication reported by Kruschke (1996a, experiment 1) involves a fictitious medical diagnosis problem, in which participants are presented with a set of symptoms and have to make a diagnosis regarding which of four diseases the 'patient' is thought to have. While the experiment (which will be detailed in a later chapter) involves four disease classifications the effect can be explained in terms of just two, a common disease, indexed C and a rare disease, indexed R. For the learning phase of the experiment $p(C) = 0.75$ and $p(R) = 0.25$. The symptoms presented are a perfect predictor of the common disease, PC, a perfect predictor of the rare disease, PR, and an irrelevant symptom, I. Only two distinct sets are presented during training. When the common disease is present the symptoms presented are the pair PC and I, when the rare disease is presented the pair is PR and I.

The inverse base-rate effect is noted in a test phase at the end of training, when participants are presented with individual symptoms and different combinations of the symptoms, and expected to make guesses regarding the most likely disease. Participants display a base-rate effect when tested with symptom I and tend to identify the symptom with the more common disease. Despite the fact that the symptom is, in fact, non-valid,

participants assign it to the class in which they have seen it most often. When presented with the pair PC+PR, participants show the inverse base-rate effect in that they tend to diagnose the *rare* disease more often than the common one.

The effect is described as 'inverse' because the base-rate would suggest that when presented with conflicting symptoms, there should be no more evidence in favour of one symptom than another. To 'break' the tie one might assume that participants would just use the base-rate information and select the common disease as more likely.

Base-rate neglect was described in section 3.2.3.1. Here the situation involves a cue which appears equally often in the context of each category such that the probabilities of each category, given the presence of the cue, are equal. One of the categories is, however, rarer than the other such that the probabilities of the cue, given each category, are different. The relations described in section 3.2.3.1, from Gluck and Bower (1988a) are shown in table 3.7.

Gluck and Bower generated test stimuli which, taken together, conformed as closely as possible to the structure given in table 3.8. One constraint was that the 'null' stimulus was not presented (i.e. a 'pattern' with none of the cues present), and this resulted in a slight difference between those probabilities given in the table and the frequencies for the experiment. This does not affect greatly the conditional probabilities of the categories given the cue and, in particular, has no effect on these conditional probabilities for cue 1, which is the cue of interest.

cue	p(cue | R)	p(cue | C)	p(R | cue)	p(C | cue)
1	0.6	0.2	0.5	0.5
2	0.4	0.3	0.31	0.69
3	0.3	0.4	0.2	0.8
4	0.2	0.6	0.14	0.86

Base-rates $p(C) = 0.75$, $p(R) = 0.25$

Table 3.7: Cue-category relationships from Gluck & Bower (1988a). R denotes rare category and C denotes the common category.

When tested with cue 1, alone, towards the end of training, participants showed a robust tendency to assign it to the rare category R. In a way this can be described as a pattern of base-rate *neglect* with respect to the cue's relationship with each category. The probability of each disease given the cue is equal and the participant may be expected to break the tie by resorting to the use of base-rate information.

As discussed in section 3.2.3.1, Gluck and Bower suggested that base-rate neglect might be explained in terms of a simple cue competition effect, used to explain blocking, which is readily modelled with a component-cue network using the LMS or Rescorla and Wagner's (1972) learning rule. Cue 1 is the best available predictor of the rare disease. It is unlikely to acquire strong associative connections with the common disease, because it is likely to occur in combination with other cues when the common category is present. When the rare category is present, it is the most likely cue to be present and, as such, will receive most of its associative weight increments in this context.

While the component cue network can account for base-rate neglect, it experiences difficulties in accounting for the inverse base-rate effect. Markman (1989) demonstrated that Gluck and Bower's (1988a) hypothesis that the component cue network might account for the inverse base-rate effect was not, in fact, correct. The absolute magnitude of the weight for the perfect predictor of the common category will grow at a faster rate than that for the perfect predictor of the rare category (Markman, 1989).

Note that the same is true of the configural-cue network. In this case the additional irrelevant cue plus perfect predictor configural nodes will be activated at exactly the same times as the perfect predictor nodes themselves. While associative strength will be split between two perfectly valid nodes, this will not make any difference to the relationship between the perfect predictors for each category, when presented together.

Kruschke's theory with regards to these base-rate effects was to suggest that both were the result of the same underlying learning principles. His intuition was that base-rates cause frequent categories to be learnt before rare ones. This learning would involve the 'commitment' of features of the common categories to predicting those categories, such that rare categories would tend to be defined in terms of whatever features were distinctive to them, or whatever features were otherwise uncommitted (Kruschke, 1996a).

In relation to the inverse base-rate effect this would suggest that the irrelevant feature is likely to be assigned to the common category prior to experience of the rarer category. When the rare category occurs, the distinctive, otherwise uncommitted feature is the perfect predictor cue. Consequently, the common category tends to be associated with the combination of the irrelevant cue and the perfect predictor of the common category. The rare category tends to just be associated with its perfect predictor.

During learning, this should result in associative strength being distributed between two cues for the common category, but being concentrated in just one for the rare category. At asymptotic levels of learning, the theory therefore predicts that if the common predictor and the rare predictor are presented together, the rare predictor will have greater associative strength and the inverse base-rate effect will be displayed.

Kruschke subsequently argued that base-rate neglect is simply an attenuated manifestation of the inverse base-rate effect. In this case the normatively irrelevant feature is infrequent for the common category and as such is not likely to be 'committed' to its prediction as much as the more commonly associated features. When the rare category occurs, this feature, being relatively uncommitted to the common category may be associated with the rare category (*ibid.*).

3.3.5.3.1. Attention to Distinctive InpuT (ADIT)

In order to formalise this theory in terms of a model, Kruschke proposed a new connectionist model called ADIT (*ibid.*). The model formalised the idea of learning involving attention to distinctive input by incorporating an algorithm to control rapid attention shifts which occurred when feedback was delivered. The model is described as a version of the component cue network and incorporates the same method of representation, i.e. a single node to represent the presence of a feature.

Each node also has an attention weight that controls the contribution of its associative strength to the decision process and also the rate of change in the associative weight. Attention weights begin each trial with a value that is, generally, a function of one over the number of cues present on a trial. Kruschke discusses alternative normalisation schemes but these are not essential to a description of the model and do not qualitatively affect its performance on the tasks referred to here.

Upon presentation of the feedback (following the measurement of the response probability) these attention weights alter, and remain normalised, in such a way as to minimise error, or network output discrepancy with the feedback vector. The operation of the algorithm is best described in terms of the inverse base-rate effect as this gives a clear understanding of why it works.

Assuming that the common category is presented first (as it will be on the majority of cases), each of the two active components will have equal attention strength but be delivering zero association strength. Altering the attention weights will not reduce error and so attention weights remain equal for the two components up to the end of the trial. At this point associative weights are altered according to the overall discrepancy, as in the Rescorla-Wagner (1972) learning rule. In ADIT the magnitude of the change is gated by the attention strength, which, for this trial will be equal for each component, so each component will get the same weight increase with respect to the correct category.

When the rare category is presented for the first time, there will be zero associative strength for its perfect predictor but the irrelevant cue will be predicting the wrong (common) category. In this case, in order to reduce overall error, the rapid attention shift algorithm must move attention away from the irrelevant cue. While the 'new' cue has a zero association weight, error is less when its attention weight is maximal than when any attention is paid to the irrelevant cue. Consequently, the irrelevant cue receives a minimal error signal whereas the distinctive feature for the rare category receives the bulk of the signal.

This pattern is repeated on each subsequent occurrence of the rare category. This tends to 'protect' the configural basis of the associative strength for the common category and locates all of the associative strength for the rare category on its perfect predictor. When testing on the conflicting cue transfer pattern, the rare category tends to win out because the cue predicting the common category has learnt in such a way as to only be delivering half of the associative strength required to reduce error to zero.

3.3.5.4. Other models incorporating rapid attention shifts

Despite its ability to account for the base-rate effects described above, ADIT clearly has a number of shortcomings in relation to its generalisability to other tasks. Most obvious of these shortcomings relates to the component cue model of representation it

employs. As discussed in previous sections representing stimuli in terms of their elemental components will not allow the model to be used to represent situations involving, for example, component-compound discriminations.

In addition, the original ADIT model does not include any perseveration of attention between learning trials. Attention weights in ADIT are reset to values determined wholly by the number of components present on a trial at the beginning of each trial. Coupled with the basic nature of the representations used, this means that ADIT would be unable to generalise its attention learning across trials.

In response to these shortcomings Kruschke and his colleagues have developed two new models. In order to address the shortcomings of the component cue form of representation, Kruschke and Johansen (1999) proposed a rapid attention shift variant of ALCOVE, known as RASHNL, for Rapid Attention SHifts 'N' Learning. This model also included a mechanism for allowing the preservation of learnt attention weights across trials. A similar mechanism is made use of in an enhanced version ADIT called EXIT, for EXtended adIT (Kruschke, in press a, Kruschke, in press b).

RASHNL is a very elaborate model, involving a number of processes absent from the basic ALCOVE model. It makes use of a variant of the attention learning algorithm used for ALCOVE (Kruschke, 1992). This algorithm is slightly different in that it 'builds-in' a certain amount of competition between dimensions in that information about all dimension weights is used in calculating the update for each individual weight. This algorithm is iterated a number of times before associative weights are updated, in order to facilitate the kind of rapid attention shifts used by ADIT.

The attention weights themselves are unbounded but a capacity limitation is implemented in a similar way to that used in ADIT, by, broadly speaking, normalising the exponentiated, unbounded weights. Attention weights are preserved from trial to trial by allowing the rapid attention shift to affect the underlying unbounded weight, then using a separate learning function at the end of the trial to preserve a proportion of the change for the start of the next trial.

The model was applied to data from experiments investigating human learning of probabilistic category structures. In order to represent the data from these experiments, RASHNL also required additional capacities such as representations of individual

dimensional salience and an 'annealing' factor which reduced the learning rates throughout the model as a function of the number of learning trials presented. These will not be discussed here as they relate to experiments which are beyond the scope of this thesis.

The RASHNL model, however, demonstrates successfully that rapid attention shifts may be applied to *dimensional* attention models as well as those which locate attention parameters on the response side of the stimulus representation. It seems likely that, given suitable settings of its 10 freely estimated parameters, RASHNL may be capable of similar tasks to its predecessor, ALCOVE.

Unfortunately, RASHNL has nothing new to say regarding the more basic shortcomings of the exemplar form of representation upon which it relies. As with other exemplar models it can only be plausibly applied to experiments where all of the stimuli have exactly the same number of components or dimensions.

Kruschke's modifications to the basic ADIT (Kruschke, 1996a) model, in the form of EXIT (Kruschke, in press a) similarly, do nothing to extend the applicability of this model beyond learning tasks which do not need any configural representations. These modifications, as mentioned above, involve giving the model the capacity to preserve attention learning between trials.

The EXIT model, the basic architecture of which is illustrated in figure 3.7, incorporates an attentional sub-system that affects the rate at which activation from basic component cue representations contributes to response strength. The attentional sub-system produces attention values in the unit range by firstly combining learnt gain parameters with direct (zero or one) activation values from the component cue representations. These latter ensure that an absent cue receives no attention.

The learnt gain contribution comes from exemplar nodes which do not have adaptive attention weights on their input dimensions. These represent each component cue's activation in terms of a location on a dimension. The location is represented as one, if the cue is present, and zero if the cue is absent. For the two cues shown in figure 3.7 this would necessitate four exemplar nodes, representing the 'points' (0,0), (0,1), (1,0) and (1,1). Each activates according to the proximity of input to the point in space which they represent.

Contribution from these exemplar nodes is gated by learnt weights. The contribution is exponentiated and multiplied by the direct contribution from the cue representations themselves. This means that the gain for an absent cue is zero, whereas the gain for a cue about which nothing has yet been learnt, is one. The gain values are then normalised using a similar function to that used for RASHNL. In this case, however, the exponentiated weighted contributions from the exemplar nodes are multiplied by the direct input from the component representations prior to the normalisation taking place. The normalised attention values are then used to gate the contribution of each cue, via its associative weight, to the decision function.

The rapid attention shifts are implemented in a similar way to those used in ADIT (Kruschke, 1996a) described above. The difference here is, firstly, that the changes are to the underlying gain values rather than to the normalised attention values. Secondly, weights from the exemplar nodes in the attention sub-system change at the end of each trial, in a way which reduces the discrepancy between the weighted output of the exemplar node and the shifted gain value. In this way attention learning may be preserved between trials. As with ADIT and RASHNL, described above, association weights are updated at the end of each trial using a variant of the Rescorla-Wagner learning rule.

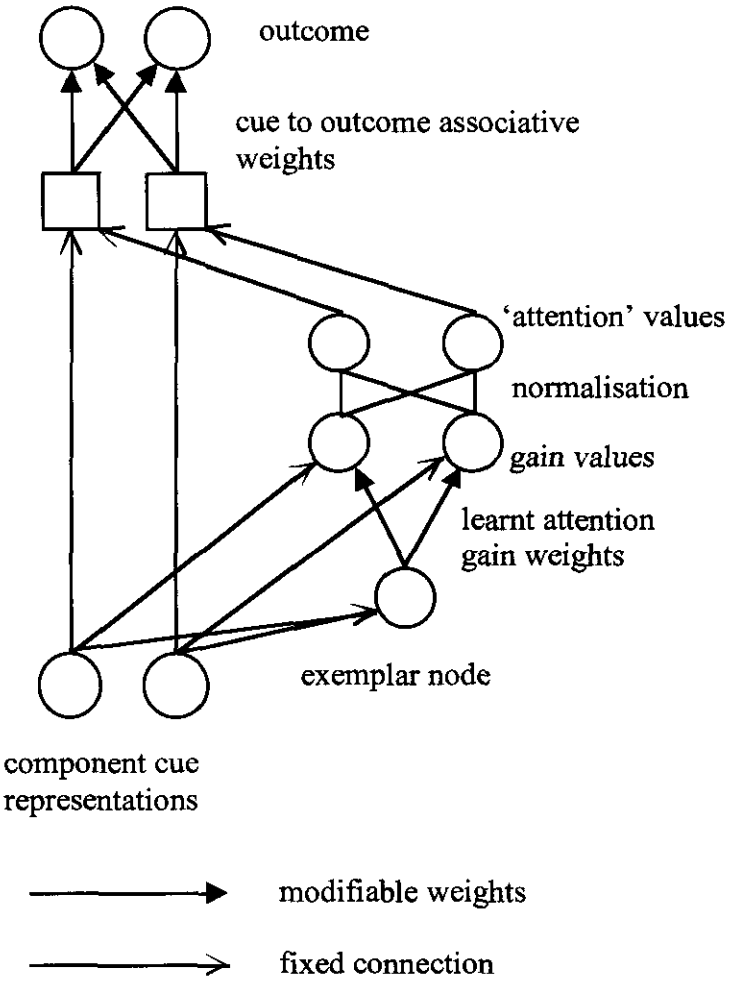

Figure 3.7: Architecture of the EXIT model (adapted from Kruschke, in press a) with two cues and two outcomes shown.

EXIT has been shown to be able to represent the same base-rate effects as ADIT (Kruschke, in press a) as well as the results from some further tests of these effects (Kruschke, in press b). In addition, Kruschke applied the model to representing data concerning blocking (Kruschke & Blair, 2000). These data suggested that blocking of learning about a novel irrelevant cue also attenuated subsequent learning about that cue relative to learning about new cues. EXIT was able to represent this effect partly because the attention weight for the novel irrelevant cue decreased during trials where it *was* irrelevant. When the cue was subsequently presented in compound with a novel cue, the

attention weight for the old cue was less than that for the new cue meaning that more associative strength accrued to the new cue (Kruschke, in press a).

EXIT is quite an ambitious model and sets out to provide a unified approach to the representation of a variety of learning effects. This unified approach is one which attempts to identify associative learning and processes of selective attention as subservient to an overall goal of 'error reduction'.

As stated above, however, its dependence on component cue representations considerably limits the applicability of the model. In addition, EXIT acquires some new problems. The exemplar nodes in the attention sub-system also make use of the somewhat questionable practice of representing the presence and absence of a cue as different locations on the same dimension. Absent cues produce net attention weights of zero, but this is only because the weighted contributions from 'null' exemplars to the gain values of absent cues are multiplied by zero prior to their use in the determination of the normalised attention vector.

As with other exemplar models, problems would be faced by EXIT when attempting to describe how exemplars are recruited to the attention sub-system when new cues and cue configurations were presented to the model. If extra cues were presented would this result in new exemplars? If so, what would their relationship be to the presences and absences of already represented cues and cue configurations?

3.4. Summary

The review of literature presented in this chapter focussed on three principal issues that ought to be taken into account when developing connectionist models of category learning. As may be apparent, the interrelatedness of these issues is such that combining the various approaches to each of them has resulted in an enormous variety of models.

This review is by no means complete. Factors such as the representation of base-rate information, effects of short-term memory processes on learning, and the representation of guessing strategies have all been identified as being the kinds of issues which may have to be taken into account when modelling category learning (e.g. Kruschke & Erickson, 1995, and Kruschke & Bradley, 1995). The way in which factors such as these interact with learning is likely to be highly dependent on the choices one makes with regards to learning rules, representation, and selective attention. As such they have not

been considered in this review but will be addressed in later chapters when they become more specifically relevant.

What appears to emerge from such a survey, is that despite the overt similarity of some of the tasks to which models have been applied, attempts to develop unitary approaches have proved fraught with difficulty. Much of this difficulty appears to reside with choices made in relation to the way in which stimuli are represented. As discussed at the end of section 3.2, models of stimulus representation have a central role in determining the way in which essential aspects of learning, such as processes of selective attention, may be modelled. Different types of representation model generalisation in markedly different ways. Any process which seeks to limit the contribution of certain aspects of a stimulus to that generalisation process, is bound to operate within the confines of what generalisation 'means' for the representation used.

To some extent, exemplar approaches such as the GCM and ALCOVE have dominated theories of category learning, precisely because they appear to offer the most coherent account of selective attention available. Where the exemplar model of representation appears to 'break down', the alternatives provided tend to focus on somewhat inflexible component cue representations.

The model of representation which would seem most capable of representing learning across the domains currently occupied by either exemplar approaches or component cue based models is the configural-cue model. As discussed, however, while the model seems to have all of the various representational possibilities which might be required by a task at its 'disposal', the lack of readily generalisable attention mechanisms considerably reduces its applicability.

The following chapters will focus on this model of representation, with the intention of proposing ways in which attentive processes may be incorporated into models making use of configural-cue based representations. The next chapter focuses on ways in which the behaviour of configural-cue networks may be modelled using approaches similar to the stochastic learning models described in section 3.1. This will be followed by chapters which represent the insights developed more specifically in the form of connectionist networks.

Chapter 4: Information transmission and learning

4.1 Information theory

Following the publication of Claude Shannon's *The Mathematical Theory of Communication* in 1948, mathematical psychology seized upon the new formalisms offered to describe communication systems. These formalisms offered ways of analysing experimental results and, more controversially, a way of describing the processes which might underlie those results (Luce, 1960).

Shannon's hugely influential work established a mathematical theory for describing the behaviour of communication systems. A schematic diagram of general communication system is shown in figure 4.1. The kind of system envisaged by Shannon is one which transmits a message from one place, its point of origin or the information source, to another, the destination.

Between these two places lay the transmitter; some means by which the message is converted into a signal suitable for transmission across a channel. The channel is simply the medium of transmission across which a signal travels to a receiver. The receiver is, for the purposes of the applications Shannon was considering, a device for performing the reverse operation to that carried out by the transmitter. In this way it 'reconstructs' the message from the signal (Shannon & Weaver, 1961, p.33-34).

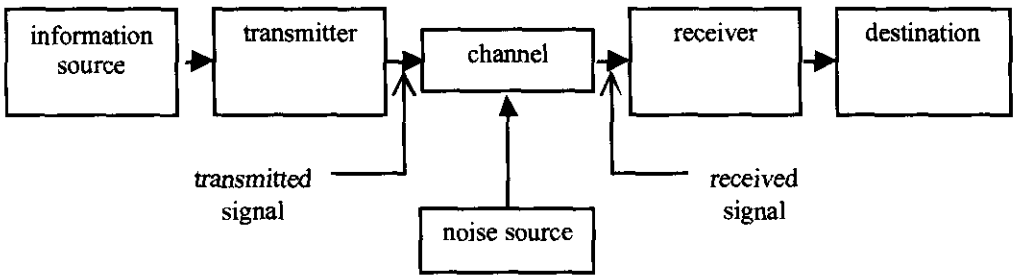

Figure 4.1: A general communication system.

The signal can be perturbed either in the channel or at either end of it (or at all points) by noise or distortion. Distortion is generally distinguished from noise in terms of the systematicity of its effects on the signal. Distortion may be described in terms of some operation which systematically alters the characteristics of the signal. If the operation

which describes the effects of distortion is known its effects can be removed by simply applying the inverse operation at the receiver or within the channel (Raisbeck, 1963, p.3). Noise, in contrast, involves statistically unpredictable perturbations, the effects of which cannot always be removed.

Shannon's theories represent two things. Firstly, a way of describing the efficiency or effectiveness of a communication system using statistical analyses of its inputs and outputs. Secondly, they provide way of defining the statistical characteristics of messages in relation to the requirements of systems capable of transmitting them.

4.1.1.Information theoretic data analysis

Shannon's basic approach to providing a measure of the effectiveness of a communication system was to examine the problem in terms of how predictable the input of the system was, given knowledge of its output. More specifically, the effectiveness of the channel was measured in terms of how much 'uncertainty' on the part of an observer, regarding the input to a channel, could be removed by observing the output.

Figure 4.2 shows a simple example of a communication system with two possible inputs and two possible outputs. The assumption is that when an input occurs, only one input will occur and it will be followed by only one output, so $\Sigma_i p(i)= \Sigma_j p(j) =1$. The diagram shows three different representations of the channel illustrating the three different probability measures which may be used to derive the effectiveness of the channel. These measures are related to one another in that, given knowledge of the joint probabilities p(i,j),

$$p(j\,|\,i)=\frac{p(i,j)}{\sum_j p(i,j)} \qquad (4.1),$$

and, via Bayes theorem,

$$p(i\,|\,j)=\frac{p(j\,|\,i)p(i)}{\sum_i p(j\,|\,i)p(i)} \qquad (4.2).$$

For the basic measure of average uncertainty, H, regarding, say, which input in figure 4.2 will occur, Shannon used the following measure,

$$H(input)=-\sum_i p(i)\log_2 p(i) \qquad (4.3).$$

H(input) is referred to as the *prior* uncertainty or entropy of the input as it relates to a situation where an observer has no knowledge of the output from the system. The same function applies to the uncertainty regarding the output with no knowledge of input, or H(output). The observer may be assumed to have knowledge of previous inputs such that the probability of each may be known or, in practice, estimated from the relative frequency of the two inputs.

Shannon used the logarithmic measure for a number of reasons. Two of these that are most relevant here are; firstly, it captures the intuition that uncertainty should increase with the number of alternative (input) states or messages and, secondly, that it be maximal when those states are equiprobable. For example, if one input is far more likely than the other, the prior uncertainty regarding which input will occur next will be less than that which might arise if both were equally likely. This uncertainty would be greater if there were three equally likely alternatives than two.

The unit of measurement here is the binary digit or 'bit', given by the base of the logarithm. Where the two inputs are equally likely ($p(i_1) = p(i_2) = 0.5$)) the prior average uncertainty regarding which input will occur is one bit. This is also expressed in terms of the rate at which the input source reduces uncertainty, on the part of an observer, every time a state or symbol occurs. The average rate of the source described above, with two equiprobable input symbols, indexed by i_1 and i_2, is one bit per symbol.

A simple case where the communication system is one that is 'supposed' to reproduce exactly the message transmitted may be used as an example. In this case, the inputs i_1 and i_2 are zero and one respectively, with j_1 and j_2 being the corresponding output messages, zero and one respectively. The maximum rate at which the system may transmit is, clearly, the same as the rate of the input source. If, however, noise was having some effect on the channel such that occasionally, for example, when a one was sent a zero would be received or output, then clearly the rate of the channel would be less than that of the input source. The rate at which the occurrence of an output symbol reduced uncertainty about the input symbol would be less than the rate that would obtain from 'directly' observing the input.

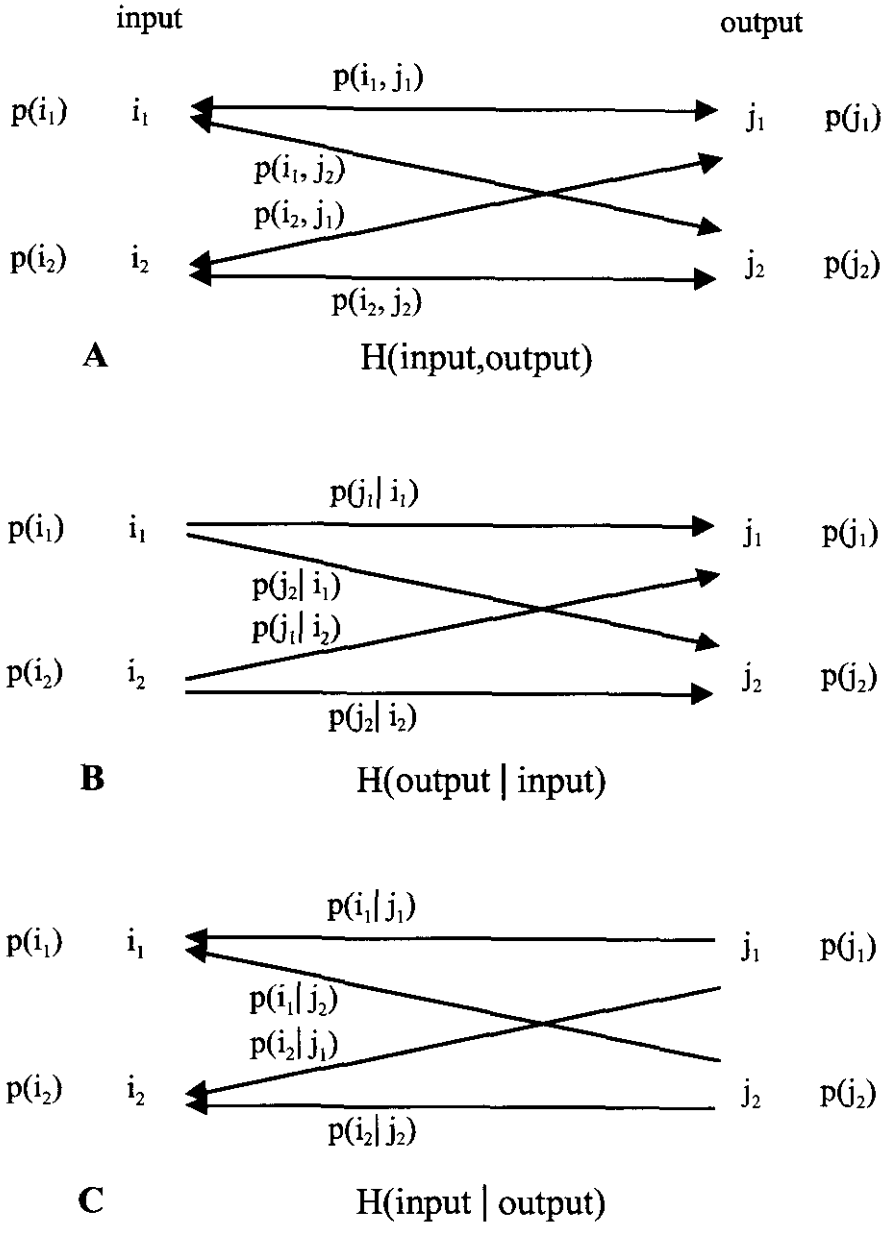

Figure 4.2: An example communication channel with possible input messages i_1 and i_2, and output messages j_1 and j_2. Probabilities shown with arrows representing A) joint probabilities, p(input,output); B) conditional probabilities of a particular output given the input, p(output | input) and C) conditional probabilities of a particular input given the output, p(input | output). Measures of uncertainty, H, are described in the text.

The measure of this loss of information from the message is known as the conditional entropy or the *equivocation* of the channel. Equivocation is the uncertainty which remains about which input symbol was sent, given knowledge of the symbol received. This measure involves the use of the conditional probabilities p(input | output) as shown in panel C of figure 4.2. It may be evaluated according to the following function,

$$H(input \mid output) = \sum_{j}\sum_{i} -p(j)p(i \mid j)\log_2 p(i \mid j) \qquad (4.4).$$

Following this the average rate of transmission between the input and the output, known as T(input; output), may be calculated in terms of the rate of the input source minus the loss of information in transmission, or equivocation,

$$T(input;output) = H(input) - H(input \mid output) \qquad (4.5).$$

Broadly speaking, the transmission rate function given above appears to provide a measure of the extent to which the input of a communication system can be predicted from knowledge of its output. The opposite scenario, whereby the average extent to which the output may be predicted from knowledge of the input is illustrated by figure 4.2 B. In this case the conditional probabilities of output given input are required. These are used to determine the average amount of uncertainty regarding the output of the system that remains, given knowledge of the input. This is known as the *ambiguity* of the channel (McGill & Quastler, 1955),

$$H(output \mid input) = \sum_{i}\sum_{j} -p(i)p(j \mid i)\log_2 p(j \mid i) \qquad (4.6).$$

Obviously, this average removed uncertainty may not be greater than the average uncertainty regarding the output itself, so T(input; output) is calculated as follows,

$$T(input;output) = H(output) - H(output \mid input) \qquad (4.7).$$

The application of these functions is subject to two assumptions regarding the processes on which variation of the input and output sources depend. The first of these is that the variation in the sources depends on *stationary* processes. Briefly, a stationary process is one whose variation is not dependent on the time across which the variation is measured. In the above example the probabilities p(i) and p(j), and the conditional probabilities p(i| j) should be the same regardless of how long the system is observed for.

The second assumption is that the processes are *ergodic*. This means that if one makes a number of sets of observations of the various sources involved, there should be no

difference in the various probabilities observed. A process then can be said to be stationary and ergodic if the process is statistically invariant with respect to the duration of observation *and* with respect to the particular record of an observation (Karbowiak, 1969, p.72).

Stationary ergodic processes are not normally observed in nature, but quasi-stationary processes, which are statistically invariant across a '*time scale* of interest', may be (*ibid.*). Furthermore, many observable processes may be said to approximate ergodic processes and thus be said to be quasi-ergodic, particularly if the records are of long duration. The requirement for both of these assumptions to hold for the processes under examination means that, for practical purposes where relative frequencies are used instead of probabilities, the above measures represent approximations to some 'true' figure. The approximation becomes increasingly reliable as the requirements of ergodicity and stationarity are approached.

These assumptions are similar to those which exist in other statistical formulations where, for example, sample size is a factor in assessing the reliability or generalisability of an index of correlation. The similarity of the theorems of communication theory to those of statistical tests such as analysis of variance prompted a number of researches to investigate the utility of the tools of communication theory in analysing experimental data from the behavioural sciences (e.g. McGill, 1954). Work was also carried out relating this method of analysis to variance analysis and correlational analysis (e.g. McGill, 1955) revealing formal parallels between the systems.

The use of the tools of communication theory to analyse experimental data treats the experimental participant as, in effect, a communication channel. For a learning experiment, for example, one would be regarding the stimuli as input sources or messages and the responses produced as the output of the communication system.

Assuming experimenter control over the stimuli, one could regard the process determining which stimuli are presented as quasi-stationary. This would be the case if the stimuli had controlled frequencies across the duration of the experiment. The minimum time-scale for which the process may be said to be stationary would be established if the presentations were organised into blocks of trials where the frequency within a block was the same as the frequency across any number of entire blocks. Obviously there would be

some deviations if the measure of relative frequency was taken from less than the number of trials in a block, or was from a number of trials not divisible without remainder by the number of trials in a block.

The responses of the participant may also be described in this way if one can assume that there was no bias on the part of the participant towards one response over another or, rather, that one can assume that this bias would not change across the duration of the experiment. Quasi-ergodicity may be similarly assumed given the above criteria. This may be extended to cover the output of a number of participants. This would be easier if there was little difference in the response biases of the participants and, preferably, no response biases at all.

A specific example might be a forced-choice identification experiment where, for example, one was training the participant to press button A, if exposed to stimulus 1, and press button B if exposed to a similar stimulus 2, where stimuli were equally likely. One could expect (assuming no response bias) the relative frequencies of stimuli and responses to remain unchanged throughout the course of the experiment. The ability of the participant to discriminate between the two stimuli, however, could be inferred from an analysis of the average frequencies with which each response occurred given each stimulus. If the equivocation or conditional entropy, estimated by representing these frequencies as conditional probabilities, was less than the average entropy of the input then, by equation 4.6, some relationship between the stimuli and the responses may be proposed.

Traditionally, these average frequencies are expressed in terms of the joint relative frequencies of stimulus-response events measured across blocks of trials and/ or between participants. This method yields a contingency table or confusion matrix with joint relative frequencies, shown below as joint probabilities, as follows;

$$\begin{array}{cc} & \textit{response} \\ & \begin{array}{cc} j_1 & j_2 \end{array} \\ \textit{stimulus} \begin{array}{c} i_1 \\ i_2 \end{array} & \begin{bmatrix} p(i_1,j_1) & p(i_1,j_2) \\ p(i_2,j_1) & p(i_2,j_2) \end{bmatrix} \end{array}$$

This approach, corresponding to the channel representation given in figure 4.2 A, was frequently used to represent data in learning, recognition, identification and choice experiments but requires the transmission rate to be calculated in a slightly different way,

$$T(input;output) = H(input) + H(output) - H(input,output) \quad (4.8).$$

Where H(input,output) is the average joint entropy of input-output events determined as follows,

$$H(input,output) = -\sum_{i,j} p(i,j)\log_2 p(i,j) \quad (4.9).$$

Note that the transmission rate is sensitive to 'order' with respect to the diagonals of the confusion matrix, irrespective of direction. If the participant always responded with j_1 to $i_1$and j_2 to i_2 then the value of T(input; output) would be 1. The same value would result if the participant always responded with j_2 to $i_1$and j_1 to i_2. If, however, the participant simply always responded with j_1, regardless of input, the value of T(input; output) would drop to zero due to the resultant drop in the value of H(output), to zero.

4.1.2 Information theoretic analysis of supervised learning experiments

The categorisation and associative learning tasks described in the previous chapter, which are the main focus of this thesis, are all cases of *supervised* learning. The participant is exposed to a stimulus and is given some form of feedback with regards to their responding. Generally, for each stimulus presented, there is an experimenter-defined correct response.

The goal of the experiment is to examine the way in which the participant learns which response is 'correct' for each stimulus. Probabilistic category structures or probabilistic reinforcement schedules are variants of this scheme in which the feedback is only partially correlated with the participant's response to a particular stimulus. In this case the correlation is generally greater than zero otherwise no learning would be expected and, in fact, one might expect irrelevance to be learned (see section 3.3.2.1).

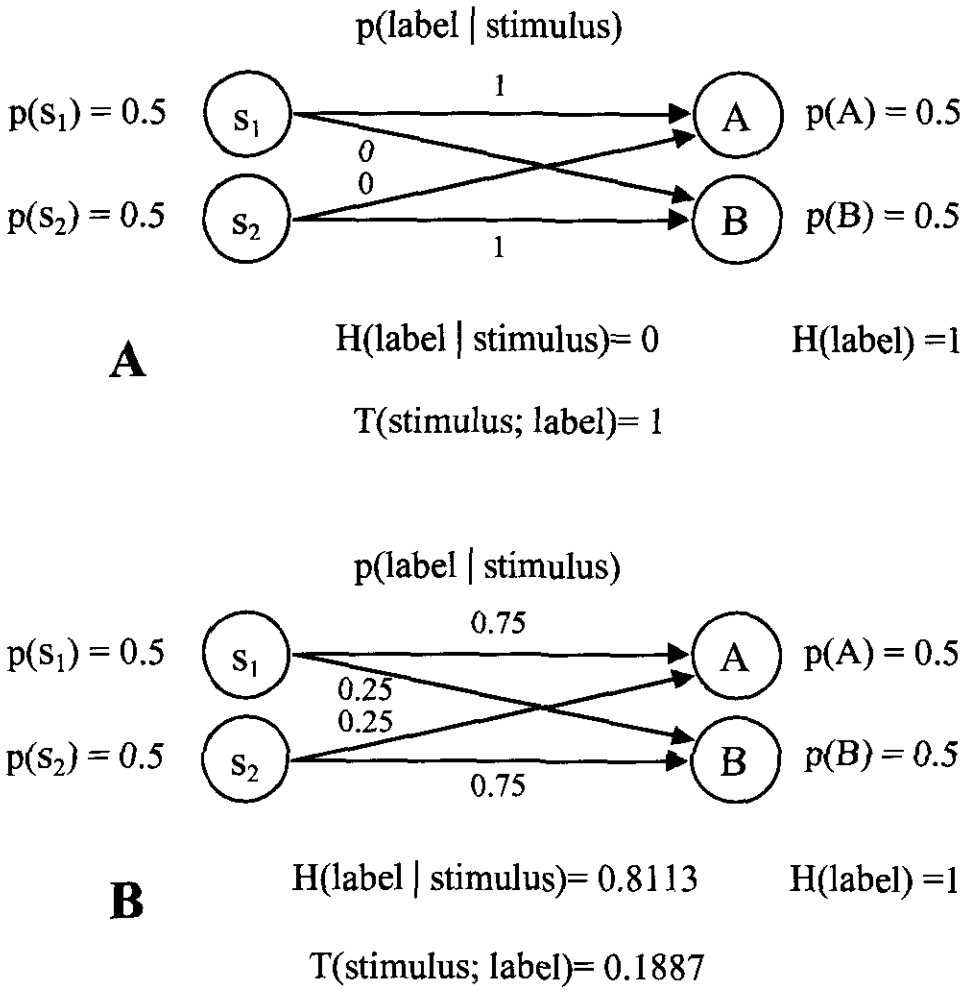

Figure 4.3: Maximum average transmission rates between stimuli and label feedback (T(stimulus; label)) for two learning tasks. Panel A shows a task where the label is perfectly predictable given the stimulus. Panel B shows a task where the label is only probabilistically related to the stimulus.

The structures of these supervised learning experiments are particularly amenable to description using the tools of communication theory. In this case, one would represent the stimuli as inputs and the *desired* responses as outputs. In a category learning experiment, for example, the stimuli are the inputs and the category labels are the outputs. Figure 4.3 shows an example of this applied to two simple learning tasks with two stimuli, s_1 and s_2, and two responses, A and B. The feedback for this experiment is assumed to be

the label and so 'transmission' may be calculated from the stimulus to the label feedback given.

In the first experiment (figure 4.3 A), the label is perfectly predictable from the stimuli. The maximum transmission rate for any experiment involving the two equiprobable labels A and B will be equal to the entropy or uncertainty regarding this source, calculated by equation 4.3 as 1 bit. The ambiguity of the channel, or the conditional uncertainty regarding the output given knowledge of the input, calculated using equation 4.6, in this case is zero. The maximum average transmission between stimuli and labels is, by equation 4.7, one bit per stimulus.

Figure 4.3 B shows a situation in which the label is only probabilistically related to the stimulus shown (given by the conditional probabilities). Using the same equations yields a maximum average transmission rate of 0.1887 bits per stimulus.

4.1.3. Multivariate signal processing analysis of Shepard *et al.*'s (1961) category structures

In the appendix of their Psychological Monograph, Shepard *et al.* (1961) suggested that the difficulty of each of their categorisation tasks could be described in terms of the number of type I classifications required to reduce uncertainty about the category label to zero. The type I category structure (see figure 2.1) requires knowledge of only one dimension to determine category membership precisely. The type II can be described as consisting of two type I structures, in that knowledge of two dimensions is required and finally type VI requires all three dimensions and therefore consists of three type I structures. The authors reasoned that the types III to V generally require between two and three dimensions and so their difficulty would be between that of types II and VI, in line with the results. The analysis is not one which describes learning, but is better expressed in terms of how many dimensions, on average, a person who knows the rule has to see before their uncertainty regarding the label is reduced to zero.

The measures they used were derived from multivariate signal processing theory (McGill, 1954, 1955). It involves calculating the difference between the basic uncertainty regarding the category label, given no knowledge of the input object, and the uncertainty

that remains regarding the label when the values of one, two, and three dimensions are known.

In the case of the Shepard *et al.* (1961) analysis, the choice of input for which uncertainty reductions were calculated was determined by an optimal, sequential progression through the dimensions. The first inputs would be the two values of one dimension; the second inputs would be the four 'points' which represent the corners of the space of the first dimension and the second dimension. Finally, the third inputs would be the eight corners of the three-dimensional space.

The values required for this analysis may be described as conditional transmission rates. For example, the measure describing how much uncertainty regarding the label is removed by knowledge of r, given that one already knows the value of q is described by the measure T(r; label | q). This measure is evaluated as follows;

$$T(r;label \mid q) = T(q,r;label) - T(q;label) \qquad (4.10).$$

To calculate T(q,r; label) one can use any of the methods given in section 4.1.1, in this case, however, the input consists of the various states (q,r) that is, the various combinations of values of variables q and r. One can work out the contribution of the final step of the sequence, T(s; label | q,r) using a similar method;

$$T(s;label \mid q,r) = T(q,r,s;label) - T(q,r;label) \qquad (4.11).$$

The basic uncertainty regarding the category label, H(label), is determined using the basic entropy equation (4.1). The probabilities of each label are 0.5 in the case of the Shepard *et al.* (1961) experiments so H(label) =1.

The values of T(input;label) for each task and each input configuration are given in table 4.1. The category structures, in relation to the dimensions, are those given in figure 2.1. Reflections and rotations would yield equivalent measures but they would apply to different dimensions.

	category structure					
input	I	II	III	IV	V	VI
q	1	0	0.18872	0.18872	0.18872	0
r	0	0	0	0.18872	0	0
s	0	0	0.18872	0.18872	0	0
q,r	1	0	0.5	0.5	0.5	0
q,s	1	1	0.5	0.5	0.5	0
r,s	0	0	0.5	0.5	0	0
q,r,s	1	1	1	1	1	1

Table 4.1: Values of T(input; label) for each input set and each of the category structures used by Shepard *et al.* (1961).

As stated above, Shepard *et al.* (1961) assumed an optimal path through the dimensions to generate a cumulative value for uncertainty reduction across the three dimensions. This would involve the conditional transmission rates being calculated for each step in this sequence to produce three values indexed as (C_1,C_2,C_3) by Shepard *et al.* For a sequence q to r to s, the C values would consist of the following measures; (T(q;label), T(r; label | q), T(s; label | q,r)). Table 4.2 shows the C values calculated for optimal paths through the dimensions. 'Sub-optimal' paths are those where more weight would be given to higher dimensional information. A sub-optimal path for the type I structure might be one where q was the last element. In which case, no uncertainty would be removed until q had been inspected; by this point all of the dimensions would have been inspected.

Shepard *et al.* (1961) reasoned that the more the participant required high dimensionality information to reduce uncertainty, the more difficult the task would be. A difficulty rating could therefore be produced by differentially weighting each of the C values and summing the products. Assuming the weight for C_1 is one, the difficulty, D, might be determined by (*ibid.* appendix),

$$D = C_1 + \alpha C_2 + \beta C_3 \qquad (4.12),$$

where α is the weight for two dimensions, and β is the weight for three-dimensional information. The authors established that in order for the values of D to be ordered in the same way as the observed task difficulty then $\beta > 1.378\alpha - 0.378$ (*ibid.*). This includes values for the weights which are equal to their dimensionality (e.g. $\alpha=2$, $\beta=3$). This set of weights is assumed for the values shown in the second to last row of the table. They yield difficulty ratings of I=1, II=2, III to V= 2.31128, and VI=3.

	category structure					
	type I	type II	type III	type IV	type V	type VI
C_1	1	0	0.1887	0.1887	0.1887	0
C_2	0	1	0.3113	0.3113	0.3113	0
C_3	0	0	0.5	0.5	0.5	1
$\Sigma_i C_i d_i$	1	2	2.3113	2.3113	2.3113	3
path(s)	qrs, qsr	qsr, sqr	qrs, qsr, sqr, srq	any	qrs, qsr	any

Table 4.2: C values and optimal paths for each of the Shepard *et al.* (1961) category structures. The second to last row gives the cumulative total when each C value is multiplied by its dimensionality, d.

4.2 Information theory and models of learning

The analysis of experimental results and structures using the tools of communication theory involves using the same functions to describe the relationship between dependent and independent variables in an experiment as one would use to describe the outputs from and inputs to a communication system. An obvious extension of this usage is to infer that the processes underlying the observable behaviour of experimental participants *are* processes of communication. For this hypothesis the connection or bond between stimuli and responses *is* a communication system and may be described using the terminology and methods of communication theory.

The representation of learning using information theoretic models is potentially problematic in terms of what aspects of the experiment the communication system is meant to represent. Traditionally, channel capacity measures have been applied to stimulus-response contingency tables in relation to such tasks as discrimination, where performance is typically measured at asymptote (see, for example Luce, 1959 and Luce, 1963 for reviews of this usage). In this case the transmission rate estimate derived from the contingency table is used to indicate the reliability with which participants may discriminate between stimuli. The implication here is that when errors are made, transmission is less than perfect. The reason for this less than perfect transmission is assumed to be noise between stimulus and response. This noise may be attributed to some limitation on the capacity of the participant to represent, uniquely, the particular stimulus being presented.

In an experiment which involves learning, the measure of transmission rate between stimulus and response changes from trial to trial. In this case it is not really possible to say what *the* transmission rate of the channel is, as the various conditional probabilities will vary from trial to trial. The communication channels being examined in the context of learning are *adaptive*.

To recap on the basics of associative learning theory, the presentation of a stimulus is followed by some decision process which, when made and implemented, is followed by some consequence for the organism. If the decision is reinforced in some way, then the assumption is that some connection between the stimulus and response is enhanced, such

that the probability of the same response following subsequent presentation of the same stimulus is increased in some way.

In communication terms, some intra-systemic representation or detector of the stimulus transmits a signal in the presence of the stimulus. This signal is received by a decision process that converts the signal, if appropriate, into an output from the system in the form of a behavioural response. Reinforcement of this response leads to a modification of the signal received by the decision process from the antecedent stimulus representation or detector, such that the signal, when subsequently transmitted, biases the decision process in favour of the reinforced response. This modification may be conceptualised as a modification to the channel between the transmitter or stimulus representation, and the receiver or decision process.

This representation of the communication between a stimulus and a response is, essentially, the basis for connectionist models. Figure 4.4 illustrates one way of mapping some of the key features of the connection between a stimulus and response onto the framework of a communication system. The modifiable bias or distortion is represented as a characteristic of the channel that contributes to the received signal. It is shown in this way as its effect is assumed to be dependent on the transmission of some signal from the stimulus representation. This signal may be regarded as, basically, neutral with respect to the decision process. This neutral signal may be described as reflecting the 'activation' of the stimulus representation. The bias, therefore, may be described as the weight on the connection.

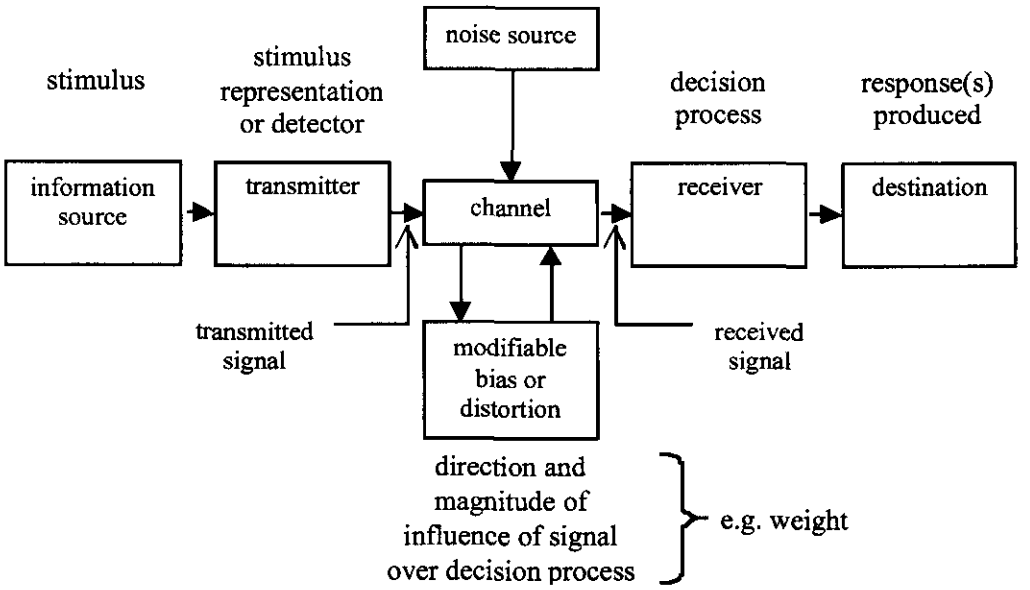

Figure 4.4: Relationship between a stimulus and response represented as a communication system with modifiable bias or distortion component to channel.

The basic assumption, in applying this form of analysis as a model of the learning process, is that a naïve participant may be described in terms of a channel with a zero initial transmission rate. Learning results in a 'modification' of this channel towards the maximum allowable by the particular structure.

With regards to the ergodicity and stationarity of the various sources involved, certain assumptions are required. As discussed above, provided that the probabilities of stimuli and outcomes, and the conditional probabilities which relate them, do not change across the course of an experiment, the experimental structure itself may be regarded as being composed of quasi-ergodic and quasi-stationary sources. As such, there is no real problem with applying the above analysis to these kinds of experiment.

For the analysis of a participant's responding, however, the relationships between stimuli and responses may not conform to these requirements. The assumptions required to make learning amenable to this form of analysis are similar to those involved in using connectionist networks to represent learning. The approach may, therefore, be most accurately described as a way of representing transmission rates in connectionist networks.

For these networks, on any given trial the probability of a particular response, given a particular stimulus, is a function of the state of the system at the time the stimulus is

presented. For that state, the conditional probabilities of each response, given each stimulus, are fixed. As such, the sources involved and their relationships can be described as ergodic and stationary and any entropies and transmission rates calculated for that particular state may be regarded as accurate measures.

When feedback is given and weights are altered, the state of the system changes. Strictly speaking, when recalculating the transmission rates, it is not the case that the maximum transmission rate for that system has changed. Rather it may be better expressed in terms of the system being a different system with a different transmission rate. Learning in connectionist networks may, in these terms, be described as producing a sequence of different systems, each with higher maximum transmission rates than the last. Generally, the assumptions that relate connectionist networks to the participants, whose behaviour in the experimental context they are meant to represent, are the same.

4.2.1. Supervised learning, feedback, and adaptive channels

When transmission measures are used to describe asymptotic responding in discrimination experiments, for example, the channel observed or inferred is between stimuli and responses. As discussed above, it is inferred that the participant knows what the correct response is for a given stimulus such that any equivocation or ambiguity measured from the contingency table may be attributable to 'noise' in the channel.

For learning, the situation is somewhat different. Increasing measures of the predictability of the response, given the stimulus, may provide an indication that learning has taken place, however a more appropriate measure of performance may be the ambiguity or conditional uncertainty of the *feedback*, given knowledge of the response. In a multidimensional categorisation task, for example, a participant may classify stimuli according to the value of one dimension, which is non-diagnostic with respect to the category label. While this makes prediction of the response, given the stimulus, easy, resulting in a high estimate of transmission between stimulus and response, the response may be wholly uncorrelated with the feedback. In this case, as far as the experiment is concerned, no learning has occurred.

As a result, the model suggested here is one in which the increase in the transmission rate of the channel, which may be inferred between stimuli and responses, is 'driven' by the uncertainty which obtains between the response produced and the feedback

which results. In other words, the increasing rate of a channel between stimuli and responses occurs *because* the particular way in which this transmission is effected decreases the conditional uncertainty of the feedback or outcome, given the response.

This assumption is, to some extent, required if one is to be able to represent learning in which the feedback signal is a simple 'right' or 'wrong' signal, rather than specific category label feedback. In this case the connection modelled cannot be between the stimulus representations and the feedback as there is no direct relationship between these two variables. As the performance of the system increases, the proportion of 'right' signals increases. This reduces the entropy of the feedback signal towards zero with the result that the transmission rate between stimuli and feedback decreases as well.

In relation to connectionist models, the above observation indicates the somewhat 'overworked' nature of a network's output nodes. In models of supervised learning the activation of output nodes is presumed to generate response probabilities. The responses having been generated, a feedback signal of some description occurs. This signal is presumed to exert an effect on the connection between stimulus representations and the decision process. The effect is proportional to the discrepancy between the response strength vector and a vector representing the desired decision.

In effect, the 'correct' decision is represented in such a way as to enable a teacher signal to affect the connection between the stimulus representations and the decision process. This requires, in relation to a communication system, the decision process or output nodes to act as receivers of response strength, generators of response probabilities, receivers for feedback signals, and 'comparators' determining the magnitude and direction of weight update signals.

One of the main problems with this approach is that learning is directional. Participants are learning to get all of the answers correct. A similar transmission rate might be measured for a channel, regardless of the nature of the feedback, if the participant was getting all of the answers wrong. For an observer of the channel, the feedback would still be perfectly predictable from knowledge of the response. In this case the observer would need to 'convert' the response using the same operation on each response to get the correct answer. The directionality, in connectionist networks is implemented by allowing the

weights to the decision process or output nodes to adapt in the direction of a perfect representation of the correct decision vector.

The assumption that learning is directional means that increases in transmission rate for the channel can be described in terms of increases in p(correct) and decreases in p(incorrect). Relative to the channels proposed here, this corresponds to two types of conditional probability. With regards to the channel between stimuli and responses the conditional probabilities representing this are p(responds A| member of A) and p(responds B | member of B) for p(correct), and p(responds A| member of B) and p(responds B | member of A) for p(incorrect).

For the channel between response and feedback this depends on whether the feedback is in the form of a category label or in the form of a right-wrong signal. For the label feedback the probabilities of interest are p(feedback A| response A) and p(feedback B | response B) corresponding to p(correct), with p(feedback A| response B) and p(feedback B | response A) corresponding to p(incorrect). For right-wrong feedback the probabilities are p(right | response A) and p (right | response B) for p(correct), and p(wrong | response A) and p (wrong | response B) for p(incorrect).

In order to model the learning process using these probabilities, it can be said that the ongoing transmission rate of the channel between stimulus representations and responses may be described in terms of T(membership; response). That is to say that the transmission rate may be described in terms of how well the response may be predicted given knowledge of which category the input stimulus was a member of. T (input; label) gives the maximum rate it may reach. Note that these are different in that T(input; label) covers the inputs to the channel, whereas T(membership; response) divides these inputs into the groups reflecting their category membership. These relationships are illustrated in figure 4.5 for a category structure with two categories and all stimuli assumed to be members of one category or the other.

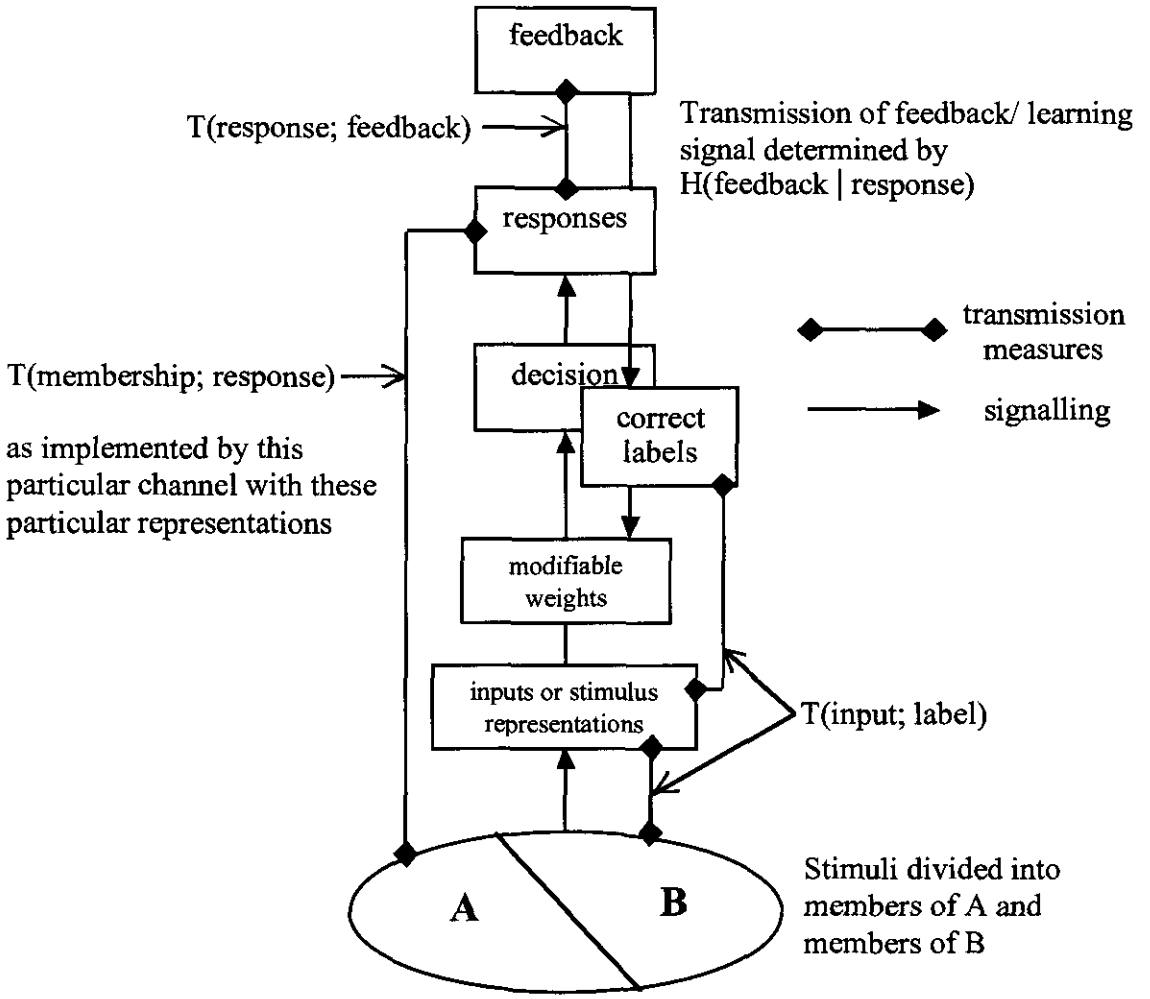

Figure 4.5: Channels and transmission relationships in a supervised learning model. Stimuli are divided into mutually exclusive sets, A and B. T(membership; response), T(response; feedback), and H(feedback | response) are changing properties of the channel in the experiment. T(input; label) is a measure fixed by the nature of the channel's input and the experimental structure.

In fact T(input; label) is an index of the extent to which the inputs to the channel may be divided into groups based on the label which applies to them. In the diagram this is represented where T(input; label) is shown as applying in both directions, towards the correct label and towards the stimuli grouped into their two categories. The adaptive channel shown is one where T(membership; response) changes as p(responds A| member of A) is altered by modification of the channel weights. Here, the member of A is the

whole stimulus. The channel in question, however, may just be getting part of that stimulus as input. The maximum transmission rate is therefore a function of how well that particular 'piece' may be used to determine category membership, or T(input; label).

4.2.2 Predicting category learning rates

The analysis in section 4.1.3 suggested that task difficulty might be described in terms of the distribution of information required to resolve the uncertainty regarding the category label. The more dimensions involved in the final rule defining membership, the harder the task will be.

The analysis of learning in relation to channel transmission rates, outlined in the previous section indicates that the values used to calculate the difficulty rating in Shepard *et al* (1961) may have a different application. The values in table 4.1 may be described as the transmission rates of channels with a particular set of inputs, e.g. values of q, qr, qrs etc., and outputs in the form of the category labels, e.g. T(input; label).

In relation to learning, this transmission rate gives the maximum, average, certainty a person may reach about the category label, given knowledge only of the particular inputs which pertain to the channel. Associative learning models typically involve the modification of some connection between a cue and a reinforced response, from an initial 'naïve' level, where prediction, given no other sources of input is at chance level. This modification proceeds towards a maximum determined by, mainly, some measure of association between the cue and reinforced responding. This may be described (in terms of channel transmission rate) as a channel increasing its rate from zero towards its maximum.

The rate also describes the *effective* rate of transmission for the signal that is resulting in this modification. This is particularly relevant for associative learning models, where teacher or error signals exert their effect according to the presence or absence of the cue. This is not dependent on any of the characteristics of the channel itself, apart from the current level of error; it is wholly a function of the correlation between a particular state of the target and the presence of an input. Its dependence on the current level of error is described in terms of the value of H(feedback | response) described in the previous section and shown in figure 4.5. This will decrease as the channel adapts in the desired direction.

The transmission rates would appear to provide an index of how rapidly, and to what extent, an associative network, given a particular set of inputs, may learn a categorisation task. The assumption is that the transmission rate of the channel is increased by increments to the conditional probability that the network produces an A as output (say), given a member of A as input. This probability will be incremented towards that which applies to the maximum transmission measure, at a *rate* that is a function of the maximum transmission measure.

An additional parameter to consider for each network would be the frequency of the individual inputs for each network. The assumption here is that transmission rate only accumulates for input when that input is present. The less frequently each input appears, the more slowly, on average, the network will learn.

Unfortunately, while the use of maximum transmission rates to predict learning rates appears almost straightforward up to this point, the approach is made more complicated by the parallel nature of the system being represented. There are a number of differences that need to be taken into account when relating what is fundamentally a sequential model of channel capacity, to a system which is using parallel channels.

4.2.2.1 Redundancy, the decision function, and the summation of parallel contributions

The multivariate signal processing approach outlined above provides an insight into the structure of the six tasks used by Shepard et al. (1961). It says nothing about the way in which the information about the various relationships may be delivered to a choice function, and used in determining which decision is made.

As discussed in the previous chapter, decision functions in learning, particularly those used in connectionist models, make use of some measure of relative evidence in favour of the various alternatives. Connectionist networks model this evidence in terms of associative strength between representations and the various alternatives in the decision process. This strength develops as a function of the correlation between the occurrence of the representation and a particular outcome.

As described above, one may describe this correlation as a measure of maximum transmission rate between the representations and the outcome. The transmission rate

measure *is*, in this sense, a measure of the associability of the representation of the input event and the outcome.

The multivariate signal processing approach suggests that information from certain channels is redundant in that, for example, knowledge of QR (i.e. knowledge of the combined values on dimensions q and r), on the type I task will provide no information that knowledge of Q alone won't have already provided. For a connectionist model in which channels between Q and QR and the outcome will be developing at the same time, this poses problems with regards to the manner in which their contributions interact.

During learning, for example, does the contribution of the QR channel provide anything extra for the decision? The decision functions used in connectionist approaches to learning suggest a particular model of the way transmissions from various simultaneously active channels may be combined to yield an overall reduction in uncertainty. In a connectionist model, contributions are summed to affect the decision process, such that redundancy may actually accelerate the rate at which uncertainty is reduced.

In order to approximate this effect using a transmission rate representation of the channels in the network, the role of these connections is proposed to be analogous to the role of redundant repetitions of a message in a sequential transmission model. Redundancy, in this respect, relates to the 'noise resistance' of the language used to transmit the message. The more times a message is repeated, the less likely it is that the receiver will be in error about it.

The idea is best understood in the sequential domain by the notion of a transmitter repeating a message over and over again. In a simple example, a set of symbols transmitted consists of ones and zeros. The job of the receiver is to determine which message has been sent. Assuming that the channel is affected by noise, one may calculate the probability of the receiver mistaking a one for a zero or a zero for a one. This is the probability of error, p(E).

The probability of error for one symbol sent is p(E), with the probability of correct reception being 1 – p(E). The probability of error for two symbols sent consecutively is $p(E)^2$, such that the probability of a receiver getting it wrong for n repetitions in a row is $p(E)^n$. The problem here for the receiver is deciding which interpretation is correct. How does it know when it is in error?

This depends on the decision function being used. A simple system may use an averaging system, whereby if there were more ones than zeros in the sequence the receiver may output a one. Using this decision scheme, its probability of error would be based on how likely erroneous averages are.

This is a function of the probability of error per symbol and the number of repetitions across which the average is taken. Using these, one can construct a binomial distribution for the probabilities of each different ratio of correct to transposed digits. The probability that the number of errors, N_E, equals r, given a total of n repetitions of the message, is given by the following,

$$p(N_E = r \mid n) = \binom{n}{r} p(E)^r (1 - p(E))^{n-r} \qquad (4.13),$$

where,

$$\binom{n}{r} = \frac{n!}{r!(n-r)!} \qquad (4.14).$$

A simple decision scheme may output a one if the sequence contains more ones than zeros. In this case calculating the probability of error is a simple matter of summing the probabilities that r is greater than n-r. Note that one may remove the possibility of a tie by using an odd sequence length. With the binomial distribution, if p(E) < 1-P(E), the average probability of error given a sequence length of n *decreases* as n increases. This decrease in average probability is exponential.

The choice functions used by parallel networks make use of redundancy in a way which has similar properties. In this case the weighted contributions of multiple nodes are summed and used to determine an overall choice probability. Provided the contributions of these nodes are in the same 'direction' they will combine to reduce error.

An important difference between parallel networks and a sequential communication channel relates to the contribution made to the decision function by uncorrelated sources. For the sequential model described above, the probability of error for each symbol received is equal. The assumption is that the error probability is a function of the channel and that each symbol is transmitted via the same channel. For a parallel network, the symbols are transmitted by different channels. If these channels were examined in isolation, their error probabilities would be different.

This examination would involve determining the probability of error, using the decision function, where the channel in question was the only channel contributing to the decision. In the case of an uncorrelated source, one whose maximum transmission rate was zero, the probability of error would be 0.5.

An important property of connectionist networks, in particular those which may have positive or negative weights, is that uncorrelated nodes, on average, have zero association weights. In this case these uncorrelated sources are not, on average, actually contributing to the decision process. In a single experiment, where the relationships between the stimuli and the outcomes do not change, the effect of individual channels on the probability of error in the decision process may be described as a function of the ongoing channel transmission rates.

In a sequential system, making use of redundancy in this way obviously increases the time required to transmit a message. It also involves a memory cost, in that the decision system must be able to store each of the n members of the sequence in order to be able to make a decision about the message transmitted at the end of the sequence. In practical communication channels this method of combating the effects of noise is not usually the most efficient solution available. Preferable solutions in these cases usually involve the use of encoding strategies that make error detection more rapid. The codes used are generally developed to maximise the efficiency of transmission for a given message source, transmitter, channel, and receiver.

It could be argued that in parallel systems such as connectionist networks, both of the costs of redundancy described above might be absent. The cost for parallel networks is limited to the overhead involved in maintaining a larger number of channels than that which may actually be absolutely necessary to transmit the message required. Some of this may be offset by the reduced requirement for high signal-to-noise ratios in individual channels.

In addition, efficient code construction requires knowledge of the problem. This in itself represents a time cost to any system which has to construct the code and transmit it at the same time. Processes analogous to re-encoding to improve efficiency may occur in cognitive systems but the difficulty remains, however, that the nature of the problem must be learnt before the relative efficiencies of different encoding strategies may be evaluated.

There may also be 'collateral' advantages to making use of parallel redundancy in that it enables graceful degradation of performance under conditions of system damage or noise increases. Related to this property, it may also facilitate representations which enable generalisation to take place.

4.2.2.1 Simplifying assumptions

It is not the intention of this section to undertake a detailed examination of the relationship between information theory and connectionist learning models. Rather the approach may be usefully deployed to allow approximate predictions of the behaviour of such models in the context of specific learning tasks. In this way, these predictions may be used to inform the design of connectionist models.

One aspect of the information-theoretic analysis, that may be helpful in this respect, is the idea that one may be able to predict rates at which components of a connectionist network may accumulate associative strength for a particular task. For the configural-cue model this is particularly apparent.

The models presented below represent simplifications of the transmission rate characteristics of the channels involved. Without these simplifications, the actual models would probably be at least as complicated as actual connectionist networks.

4.2.2.1.1 Spatial sub-channel representation

For this approach, each spatial sub-channel is assumed to progress from a naïve state or zero transmission rate, towards the maximum transmission rate allowable by the category structure. As discussed above, the conditional probabilities which are assumed to alter for this channel are p(correct) and p(incorrect) (equal to 1-p(correct)). As such regardless of the dimensionality of the channel, it is represented in terms of a binary symmetrical channel.

This simplification involves each channel being conceptualised as having two inputs and two outputs. The inputs may be described as members of A and members of B with the outputs being A and B. The maximum transmission rate, in this case, refers to the extent to which members of A may be distinguished from members of B on the basis of the cues or cue-configurations represented by the channel.

As stated, learning involves the alteration of the p(correct) (and p(incorrect)) probabilities subject to the difference between the channel's current transmission rate and

the maximum which it may achieve. For certain channels, this means that the asymptotic probabilities approached reflect the way in which the maximum transmission rate is realised in a binary symmetrical channel, rather than in a channel with, say, more than two input sources.

This is particularly the case for the partially valid two-dimensional channels in the type III to V structures. These channels have a maximum average transmission rate of 0.5. This is realised in terms of it having two sources which are perfectly valid and two sources which have no validity whatsoever.

The actual accumulation of transmission rate by these channels is likely to involve p(correct) for its valid sources approaching unity, and p(correct) for its non-valid sources remaining at 0.5. Calculating the transmission rate in this case would involve averaging the transmission rate for all of the sources in the channel.

With the binary symmetrical channel a 0.5 transmission rate is realised by p(correct) being approximately 0.88997. The simplification here is that all of the sources in a channel are assumed to have identical transmission rates that are equal to what would, in fact, be the average rate. If all four sources in the partially valid two-dimensional channels had a maximum transmission rate of 0.5, then their individual probability of being correct on a given trial would be 0.88997. The reason for introducing this averaging process is to allow the channel to be treated as a whole, rather than as a collection of individual sources.

The effect of this simplification on the learning rate for these channels is likely to be to accelerate the rate at which they reach a maximum. For these channels the source probability is 0.25 but the maximum value of p(correct) is 0.88997. In 'reality' two of the sources have no validity and two are perfectly valid. The probability required for the valid sources, to realise a 0.5 rate is unity. As such, the partially valid two-dimensional channels are likely to learn at a slightly faster relative rate than those in a connectionist network, as they can reach the 0.5 capacity by incrementing sources to a 0.88997 p(correct) value.

This is not really a factor for any of the other channels, as either they have unit validity or they already are binary symmetrical channels so the representation is not problematic.

4.2.2.1.2 Interaction of channels and the representation of directionality in learning

The second problem concerns the interaction of individual channel capacities to yield an overall transmission rate. The method chosen was to represent an overall error probability as a function of e raised to a negative power for the sum of the ongoing transmission rates for the channels. There are two problems for this representation.

Firstly it does not take into account the directionality of learning which occurs in connectionist learning schemes. This is only really a problem for the partially valid one-dimensional channels in the types III to V structures. For these channels their realisation in a network is likely to involve weights which provide a positive contribution to certainty on 0.75 of stimuli, and a *negative* contribution to certainty on the remaining 0.25 stimuli.

The representation of an input actually reducing certainty about an outcome is problematic for information theory. In this case the negative contribution to the certainty must be expressed in terms of the channel actually *increasing* the probability of error on the occasions when its weight is in the wrong direction for an output. In the models described below this negative aspect to the contribution has been ignored.

The problem with directionality is one that relates to the interaction of the individual channels in the network to produce an overall error probability. It could be argued that, in networks where learning is based on an overall error signal, the negative contribution made by these channels on certain trials may have a positive effect on the learning rates of channels which *are* valid on these trials. Because individual trial learning is not represented in these models either, these effects will go unrepresented. It may be an appropriate approximation to regard all of the contributions made by these channels as positive. For certain trials, however, the 'positivity' is an effect 'distributed' across other channels.

The second problem with regards to interaction relates to redundancy. While it was suggested in the previous section that redundancy would have a form of additive effect on overall transmission rate, the extent to which this is true depends on the task. Again, this is only really a problem for category structures III to V and relates to partially valid channels. The problem is with the interaction of the partially valid two-dimensional channels.

For example, in the type V task the two partially valid two-dimensional channels are not capable, when their outputs are combined, of predicting the category label for all

stimuli. The two channels 'overlap' and both successfully predict the two central category members. Each channels' remaining valid sources predict, alone, the membership of the peripheral members. The sources which are active in these channels when the exception members of the structure are present are non-valid. The result is a combined maximum transmission rate of 0.75. The same problem is present for certain of the relationships in the types III and IV structures. In the type III structure, all three two-dimensional channels have a combined maximum transmission rate of unity, as does a combination of QR and RS. Combinations of channels QS and RS, however, only have a maximum of 0.75. For type IV, all two-channel combinations have a maximum transmission rate of 0.75 but all three together yield a combined rate of one.

To a great extent this problem is 'invisible' for these category structures. One cannot test the model using a combination of any of these two-dimensional sources without also activating the fully valid three-dimensional channel.

Although differences are observed between the response probabilities for different stimuli, which are likely to correspond to their logical status as central, peripheral or exceptional members, this basic model is not designed to represent these differences. As stated above, in order to simplify the channel representation, the behaviour of all channels on all stimuli is represented in terms of its average performance, as such, all differences are 'averaged out'.

It is possible to use this approach to address differences between stimuli. To do this one would simply calculate the sum of transmission rates for all channels for each stimulus. The average channel transmission rates are simply the averages of individual stimulus rates (or the sum of the transmission rates for a stimulus multiplied by the probability of that stimulus). When calculating the combined maximum rates for each stimulus the result is a maximum 'redundancy' measure for each stimulus. These measures are highest for central members, with peripheral members next and lowest for exception members.

The summed transmission rate measure does, however, pose some problems for any attempt to generalise the approach to other learning problems. Perhaps most obvious would be a structure involving stimuli with two wholly redundant dimensions, neither of which is capable of perfectly predicting the membership. Two redundant channels with

maximum transmission rates of 0.5 could not, in this case, combine to yield a rate greater than 0.5.

In order to address this possibility one could introduce an asymptotic limit on the learning function for the network, which consisted of the maximum possible transmission rate given the representations available to the network. For the Shepard *et al.* (1961) tasks this limit would be unity, and would be given by the maximum transmission rate T(q,r,s;l) in table 4.1.

4.2.2.1 Representing ongoing performance and learning

As discussed above, simply summing the transmission rates for all of the contributing channels does not provide an adequate way of representing the overall performance of the multichannel network. For the first average trial, at least, each 'sub-network' of the organisation will increment its transmission rate according to the rates given in the above tables. While approximating the combined transmission rate using simple summation was used effectively by Bartos and LeVoi (2001), a measure is used here which takes into account the role of redundancy described above. In order to index the increase in overall performance, a measure is required to determine the average probability of error.

In order to represent the role of redundancy in this combined rate measure, the following simple measure was used to define the average probability of error, p(E), of the network on an average trial,

$$p(E) = 0.5e^{-g\sum_{i}\tau(i;l)} \qquad (4.15).$$

The first part of the equation, the 0.5, gives the basic probability of error without any contribution from the channels. This value is multiplied by e raised to the negative power of the summed channel contributions. Each channel's contribution is given by the transmission rate $\tau(i;l)$ which is the transmission rate for channel i to the receivers at *l*. The whole is multiplied by a gain constant, g.

As may be apparent, the nature of equation 4.15 is such that 1-p(E) will resemble the top half of a logistic, such as those described in the previous chapter in relation to choice functions. This particular form was used to represent the 'multiplicative' role of

redundancy (described in section 4.2.2.1) more explicitly, although a logistic is likely to produce equivalent results.

The transmission rate measure for the channel is based on a binary symmetrical channel with two probabilities being altered by learning. As discussed above, the variables being altered are the conditional probabilities p(outputs label A | input is a member of category A), known as p(correct), and p(outputs label B | input is a member of label A), known as p(incorrect). Representation of the B channels is not necessary here, as their development will be symmetrical and subject to identical constraints. These probabilities are assumed to begin at 0.5, with the 'ongoing' or current transmission rate of a channel between input set i and the labels l, called $\tau(i; l)$, given by the following,

$$\tau(i;l) = 1 + \left(\left(p(correct)\log_2 p(correct)\right) + \left(p(incorrect)\log_2 p(incorrect)\right)\right) \quad (4.16).$$

The most obvious correspondence between the set of networks given above and an actual parallel network model is Gluck and Bower's (1988b) configural-cue model, described in the previous chapter. The configural-cue model also employs a least mean squares learning algorithm, similar to Rescorla & Wagner's (1972) rule, which produces error signals proportional to the total error on a given trial. This can be implemented, to some extent, by using the transmission rate update approach to produce average trial-by-trial performance curves. As discussed in section 4.2.2.1, however, it is not possible to represent the interactive nature of this learning rule due to the 'averaging out' involved in the channel representations.

The increment in this model will, like the network model, be controlled by the total 'lack' of transmission rate. While transmission rates will head towards the entropy of the labels, H(label), the maximum attainable will not just be this measure. The asymptotic total transmission rate will be the maximum that can be achieved by this particular organisation. This level, in the case of the three-dimensional tasks described here will be equal to T(q,r,s; label). This value is given in table 4.1 and is 1 in all cases.

For these tasks the category label is predictable given knowledge of, at most, all three dimensions. If one was representing a task where this knowledge was not always enough to enable perfect prediction of the category label, then the maximum transmission rate attainable would be less than the entropy of the label. In effect this measure may be

used to say whether a task is linearly separable, given the ability to represent the inputs at a given dimensionality. All of these tasks are separable using, at most, three dimensions.

The change in p(correct) is, as described above, governed by overall lack of rate, frequency of the representations in each channel, the maximum transmission rates, and a learning rate constant (k).

$$\Delta p(correct)_i = \left(T(q,r,s;l) - \sum_i \tau(total;l) \right) \left(T(i;l) - \tau(i;l)\right) \frac{1}{n_i} k \qquad (4.17),$$

T(i; *l*) refers to the maximum rates shown in table 4.1. Its use in the update equation is to limit the rate of a channel to its maximum, given the task. On an average trial, a channel will not be able to increase its transmission rate beyond that of its maximum average rate. The term n_i is the number of input sources for the channel, 2 for a one-dimensional, 4 for a two-dimensional, and 8 for the three-dimensional channels. The transmission rate τ(total; *l*) is the combined transmission rate. It is, in a sense, equivalent to an ongoing version of the transmission rate T(response;feedback) shown in figure 4.5. It is calculated from the error probability in equation 4.15, p(E), as follows,

$$\tau(total;l) = 1 + \left(\left(p(E) \log_2 p(E) \right) + \left(\left(1 - p(E) \right) \log_2 \left(1 - p(E) \right) \right) \right) \qquad (4.18).$$

Figure 4.6 illustrates the learning curves, in terms of 1-p(E), produced for the six tasks using the above equations and a value of 0.05 for k and 4 for the gain parameter g. The graph shows the block-by-block accumulation of total transmission rate. The block is an average of 16 trials.

The model correctly predicts that the type I is easier than the type II, which in turn is easier than the type VI task. As can be seen there is almost no difference in performance on types II, III and IV tasks, with type V being slightly more difficult. This ordering is somewhat consistent with the performance described by Gluck and Bower (1988b, p.188) in relation to their basic configural-cue network. Relative performance on the six category structures for the configural-cue network changes across the course of learning. Initially type II is more difficult than types III and IV but marginally easier than type V. As learning progresses performance on the type II task overtakes that on the type IV task, but remains worse than the type II task (*ibid.*).

These graphs do not capture the change in performance that appears to give the type II a late advantage over the type IV for the configural-cue model. This is unsurprising,

as the model never actually deals with the specific objects. For the 'real' configural-cue network, the differences between objects become important, particularly late on in learning. As discussed in chapter 2, the objects in types III to V have different statuses in relation to the number of separate rules that apply to them. Rescorla & Wagner's (1972) rule will operate on these differences, as performance nears asymptote, by producing negative teacher signals when overall output is greater than one. The more distributed the rules are and the more their application varies from object to object, the more this effect will come into play.

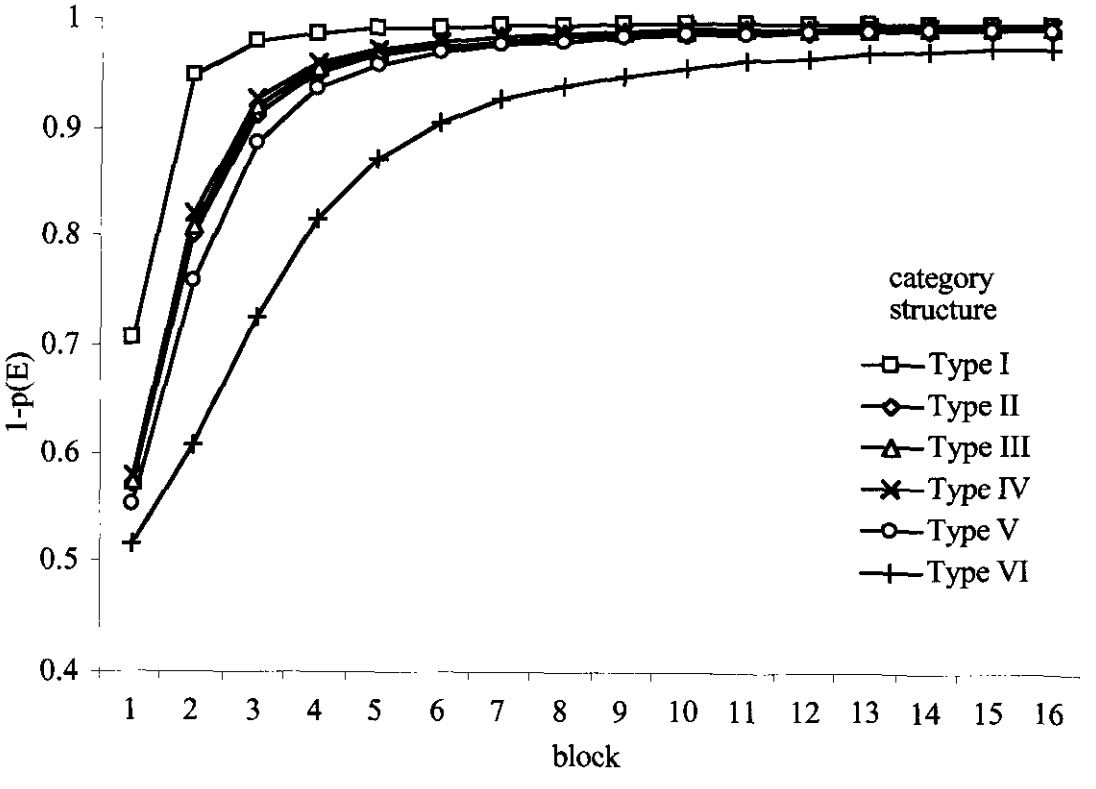

Figure 4.6: Increase in 1-p(E) by the transmission rate model of the configural-cue organisation across 16 blocks of 16 'average' trials for each of the six category structures from Shepard *et al.* (1961).

As figure 4.7 shows, the type IV structure results in a very distributed pattern of transmission rates. Because the transmission rate model does not deal with specific objects, it does not reflect the fact that the two-dimensional rules are likely to be instantiated at different times, only all active together for the 'prototype' members of each category.

Using this parallel transmission rate approach highlights important differences between Shepard *et al.*'s (1961) use of the approach to describe the complexity of a 'minimal' rule, and its use in determining how that rule may have been developed in the first place. The parallel approach requires that all of the information is, to some extent, available to the decision and learning process at the same time. The optimal path through the dimensions assumed does not apply to a parallel approach, where the general idea is that all 'paths' will be occurring at the same time.

While the *optimal* distribution of information may offer an index of difficulty, there is no mechanism specified to suggest how that optimal distribution is realised. The use of different parameters applied to different dimensionality channels, while in line with Shepard *et al.*'s (1961) fitting of the distribution to the difficulty, is not sufficient to explain the differences in difficulty. The frequency parameters employed here are analogous to these dimensionality parameters, attenuating learning according to the dimensionality, but they do not result in the required difference. Application of this type of parameter to the configural-cue model, on top of the 'natural' lower frequency of higher dimensionality representations, does not substantially improve the fit of the model to the data (Nosofsky *et al.*, 1994).

As previously discussed, the requirement seems to be for some form of selective attention process. This is envisaged as making use of higher order aspects of the information, and/ or employing architectural constraints to enhance the learning and/ or contribution of lower dimensionality, high validity rules relative to high dimensional and/ or low validity rules.

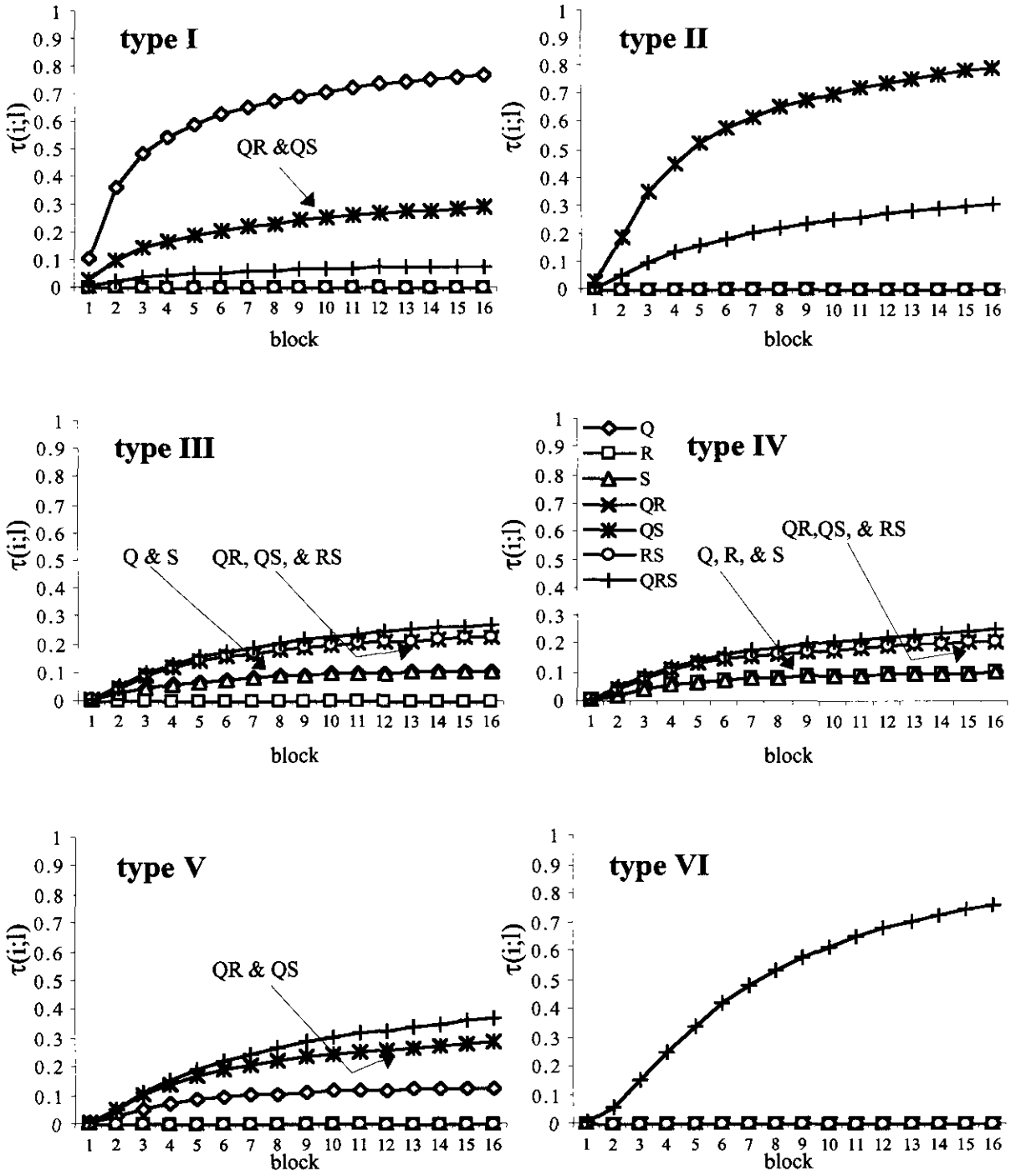

Figure 4.7: Individual channel transmission rates from the configural-cue organisation across 16 blocks of 16 'average' trials for each of the six category structures from Shepard *et al.* (1961). Key shown for type IV applies to all graphs.

4.3 Selective attention and channel transmission rates

The previous chapter discussed selective attention as a process which controls the associability of cues. Several approaches were discussed and the differences between the ways in which the associability of cues could be controlled were contrasted. The simplest approach, in the form of Rescorla and Wagner's (1972) learning model, suggested that cue associability be attenuated according to how well the outcome is already predicted by other available cues. While being extremely influential, the rule cannot explain alone, regardless of the representations used, the differences in difficulty for the Shepard *et al.* (1961) tasks.

Other approaches suggested a secondary process, represented by a set modifiable weights which might vary according to the relevance, or relative relevance, of particular aspects of the stimuli. These weights generally exert their influence according to two distinct methods.

One method is to use these weights to control the contribution made by certain representations to the decision process. The weights, in this case, which generally also control associative learning rates, act as a gate on the output and learning functions of, either individual representations, or sets of representations. Modular approaches and models such as ADIT (Kruschke, 1996a) make use of this method.

The second method is exemplified by dimensional attention models such as ALCOVE (Kruschke, 1992). Here dimensional attention weights control the contribution of certain parts of the stimulus towards the activation of representations. As activation generally controls both the contribution made by a representation and its associative learning rate, this method achieves similar effects to the first method described above.

The intuitive difference between these two approaches, in relation to transmission rates, may be expressed in terms of the parameters affected by the algorithms. In the case of those rules where the extra weight does not affect activation, it may be suggested that it is the capacity of the channel ($\tau(i; I)$ for example) which is being directly multiplied. For ALCOVE it would appear that it might be the conditional probabilities that underlie transmission rate, which are being altered. This role is subtle in ALCOVE, however, in that because it is also dependent on the recruitment of exemplars to represent lower dimensionality information, the algorithm also alters the *number* of channels representing relevant dimensional information.

The following outlines three approaches to selective attention, which augment the basic transmission rate implementation of the configural-cue model. Two of these models locate the cue associability parameter on the target side of the representations, and the other assigns it to the dimensions underlying each representation.

4.3.1. Transmission rate squared: independent channel associability weights

As discussed in the previous chapter, the configural-cue network fails to predict the exact order of difficulty of the Shepard *et al.* (1961) tasks. The reason for this is that it appears to give too much weight to the multiple, fully valid two-dimensional cues in tasks III to V. For at least types III and IV, there are more fully valid cue configurations than there are for the type II. Each learns according to its own frequency and validity and consequently learning is faster for the types III and IV than it is for the type II.

Nosofsky *et al.* (1994), suggested that performance of the basic configural-cue model might be improved if the validity of the space in which representations occurred was taken into account. As discussed in the last chapter, these authors attempted to represent this using the *dimensionalized* adaptive learning rate (DALR) model (see section 3.3.4.1.1). Although the approach failed to predict the relative difficulties of the tasks, more appropriate measures and uses of average capacity may enable such an organisation to succeed in these tasks.

The transmission rate approach includes ready-made average performance parameters for each module, and it is actually surprisingly simple to produce a qualitative fit to the Shepard *et al.* (1961) and Nosofsky *et al.* (1994) data by making use of them. The assumption is that the ongoing transmission rate of each module or sub-network is multiplicatively gated by a value that reflects this ongoing rate. In simple terms, if the ongoing rate for each channel is squared prior to summing in the error calculation, then the basic model above can be made to predict the order of difficulty. This requires an alteration to equation 4.15 to reflect that the total rate is now a function of the sum of squared rates as follows,

$$p(E) = 0.5e^{-g\sum_{i}\tau(i:I)^2} \qquad (4.18).$$

Figure 4.8 shows the accumulation of summed squared transmission rates by the model, with k = 0.075 and g=4. As can be seen the type II is now learnt more quickly than the types III to V. The squaring of capacities obviously affects the contribution of the partially valid channels in the type III to V structures. This effect will not be so great on the fully diagnostic channels of the type II structure (or anywhere else they appear), and as such type II learning is enhanced relative to learning of types III to V. Early performance appears to be attenuated slightly, resulting in a sigmoidal shape for the curves.

Comparing figures 4.9 and 4.7 also shows that all of the channels have to reach higher individual rates before error is reduced to zero. This again will affect types III and IV more than type II as, apart from the three-dimensional channel, all of the channels in the rule-plus-exception structures have lower learning rates due to their lower maximum transmission rates.

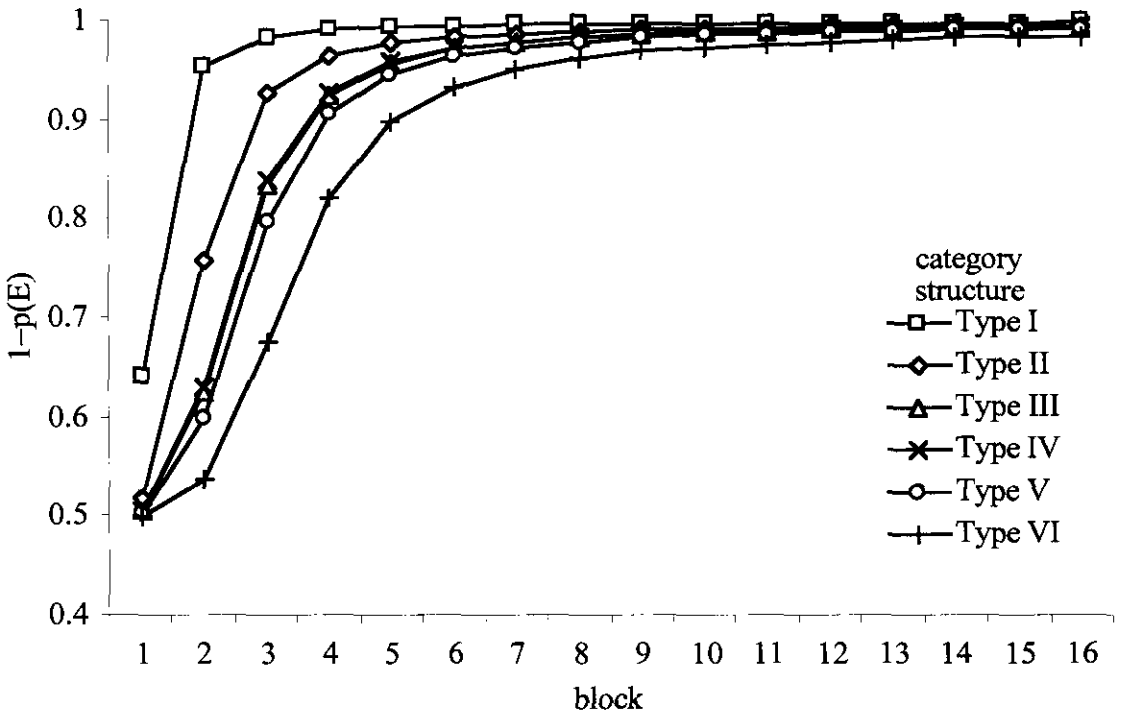

Figure 4.8: Increase in 1-p(E) by the configural-cue organisation with 'output' rate squared prior to summing across 16 blocks of 16 'average' trials for each of the six category structures from Shepard *et al.* (1961).

It is interesting to note that this model does not imply that learning rate is squared or multiplied by average transmission rate. While the measure indirectly affects learning by enhancing the overall error signal in relation to the channel's discrepancy from its maximum, the update rule is fundamentally the same.

In terms of implementation using a connectionist framework, this model does not simply imply that the output from each representation, or even channel, is squared. As will be detailed in the next chapter, the requirement is for a separate weight for each channel that tracks the *average* performance of the channel as a whole. The principal reason for this is the characteristics of the semi-valid two-dimensional channels in task types III to V.

As discussed above, these channels contain two representations that are perfectly valid, and two that are non-valid. Squaring the output from these channels will not affect the status of these individual sources in any way which is different to the way it affects the fully valid two-dimensional channel in the type II task. As with the plain configural-cue model, there are more fully valid sources in types III and IV than in the type II structure; as a result one would expect little difference in relative task difficulty for the model.

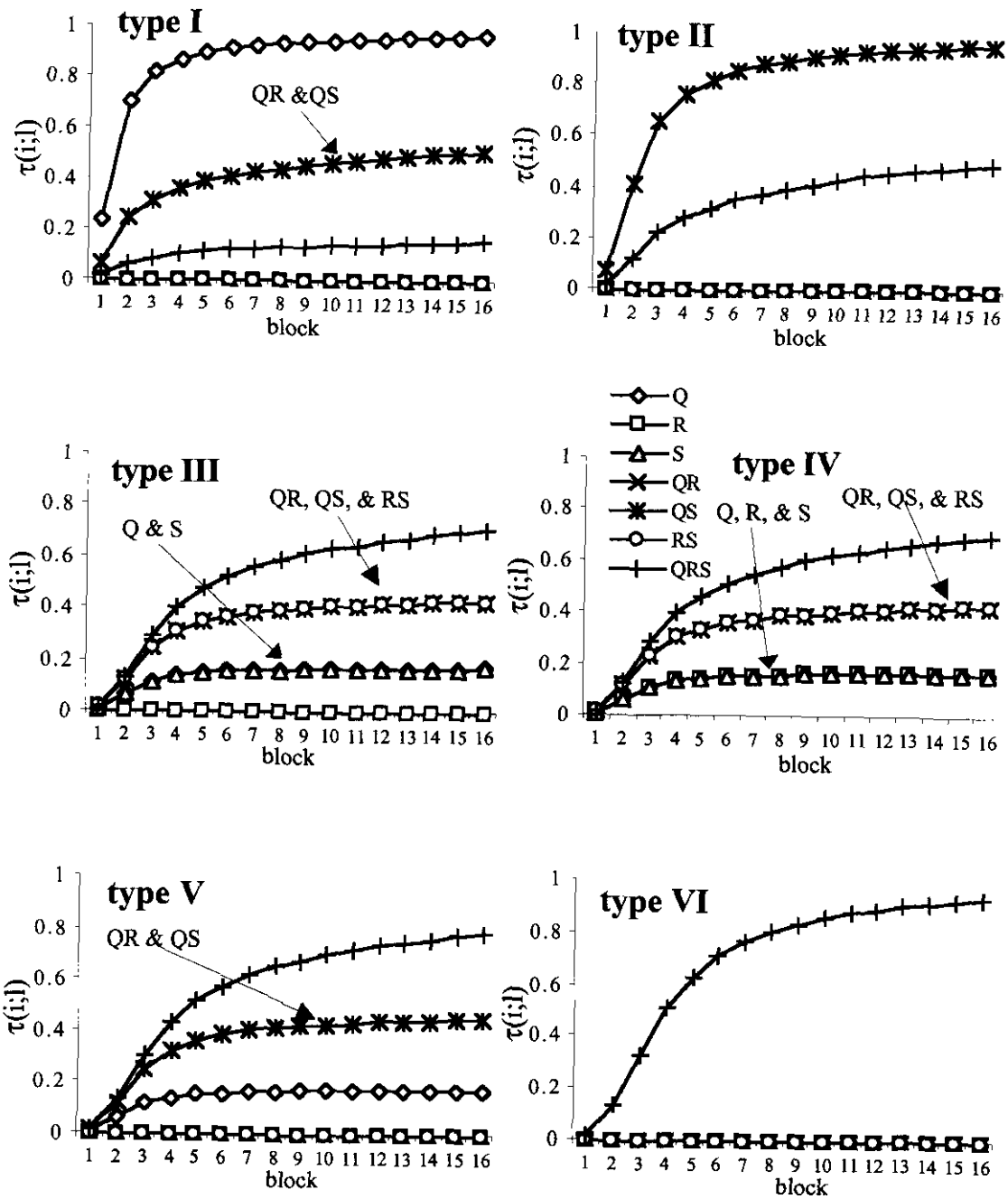

Figure 4.9: Individual channel capacities from the 'transmission rate squared' organisation across 16 blocks of 16 'average' trials for each of the six category structures from Shepard *et al.* (1961). Key shown for type IV applies to all graphs.

4.3.2 Relative channel validity and associability weights

The Mackintosh (1975) conception of cue associability weights implied that these weights would apply to each cue, and evolve according to some comparison of the effectiveness of the cue, relative to that of other cues present on the same trial.

The next model attempts to represent Mackintosh's (1975) concepts of weights which represent relative validity using *modular* associability weights. Because the channel capacities describe the average characteristics of the spatial array, all representations within a spatial channel have equal status to one another. Locating another set of weights on the spatial channels will enable their values to be determined by simply comparing the average characteristics of channels.

Unlike the 'transmission rates squared' approach described above, this will require one to specifically model the evolution of the association weights. The resultant model is 'half connectionist' in that while the individual association weights between the representations and the category label are modelled in terms of the average characteristics of their channel, the second set of weights are modelled specifically.

These weights may undergo positive or negative changes. In this model they are represented as weights which can take on positive or negative values, but mediate associability according to a logistic transform of the weight value. Conceptual difficulties regarding negative associability are consequently avoided as 'associability' of a channel varies between zero and one.

Total transmission rate, for a particular iteration, is derived from the error probability as given for the previous models. In this model the error probability is calculated in a similar way except, for the multiplication of each channel transmission rate by the associability weight, a, for the channel,

$$p(E) = 0.5e^{-g\sum_i \tau(i;l)a_i} \qquad (4.19).$$

The weights also affect the learning in that updates are multiplied by the associability weight of the channel as follows,

$$\Delta p(correct)_i = \big(T(q,r,s;l) - \tau(total;l)\big)\big(T(i;l) - \tau(i;l)\big)a_i \frac{1}{n_i} k \qquad (4.20).$$

The associability of the channel is given by the parameter a_i, this is derived from the associability weight α_i as follows;

$$a_i = \frac{1}{1+e^{-\alpha_i}} \qquad (4.21).$$

Updates to the associability weights are determined by subtracting from the transmission rate of the channel, the transmission rate of all of the channels. The sum of these differences determines the direction and magnitude of the change. Change to associability weight α_i is calculated after each trial as follows,

$$\Delta\alpha_i = \left(T(q,r,s;l) - \tau(total;l)\right)k_\alpha \sum_{j\neq i} \tau(i;l) - \tau(j;l) \qquad (4.22),$$

where k_α is a fixed update rate. These weights are initialised at zero allowing an initial associability for all channels of 0.5 via equation 4.21.

Figure 4.10 shows the accumulation of total transmission rate by the model across 16 blocks of 16 trials using k=0.075, g=4, and k_α =0.15. While similar to figure 4.8, showing the experimentally observed order of difficulty, there are slight differences. The grouping of types III to V is perhaps more marked with the type V structure actually overtaking the types III and IV structures. The reason for this is that there are larger numbers of channels with some validity for types III and IV, than for the type V. The associability weights are, therefore, likely to change less for these two structures. The more channels there are contributing, the less difference there will be between them.

Figures 4.11 and 4.12 illustrate this pattern quite clearly. Comparing figure 4.11 with 4.7 and 4.9 shows that the transmission rates reached are about midway between those seen in the conventional configural-cue organisation and the transmission rate squared version. The associabilities shown in figure 4.12 mediate an approximate 'squaring' of output capacities. While they are not equal to the capacities they are ordered in a similar way. The associabilities for valid channels in the type V structure grow to a larger size than those in the types III and IV structures due to the relative absence of competition.

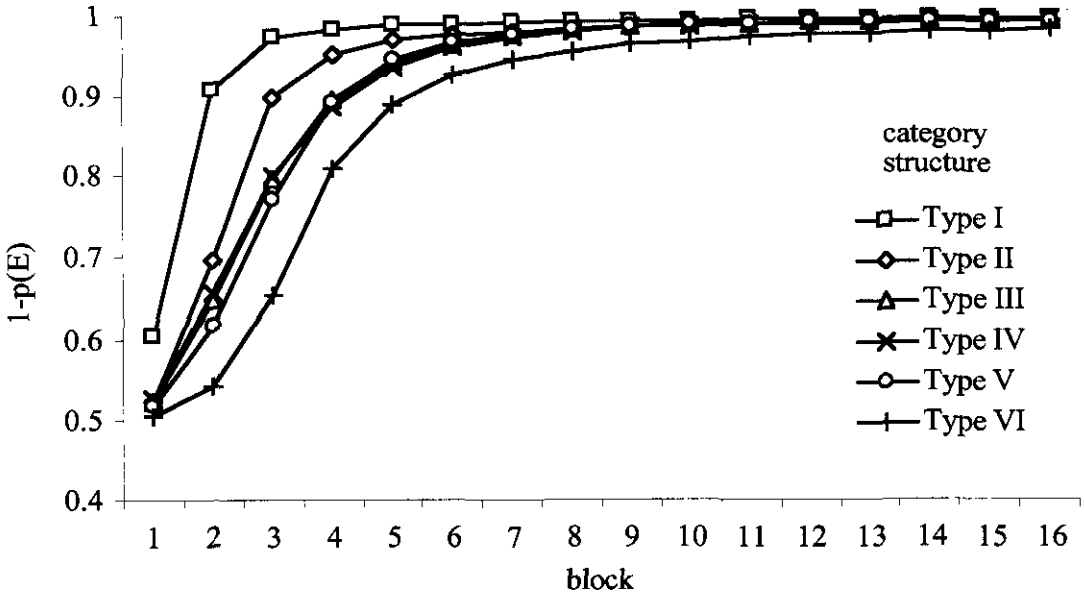

Figure 4.10: Increase in 1-p(E) by the configural-cue organisation with channel associability weights across 16 blocks of 16 'average' trials for each of the six category structures from Shepard *et al.* (1961).

This represents something of a conceptual shortcoming in the model in that the indication is that the more irrelevant dimensions there are in a task, the faster the relevant channels' weights are likely to increase. This issue will be addressed in the next chapter when connectionist implementations of these models are developed.

Another difference between this model and the previous one is the level of contribution or transmission reached by the redundant valid channels in the types I and II structures (qr, qs, and qrs for the type I, and qrs for the type II). The associability weights for these channels either increase very slowly from a zero value (0.5 value of a_i) or drop such that the net effect on output will be less than for a transmission rate squared approach. In addition the role of the associability parameters in learning curtails the development of channels with relatively lower validity. With their higher numbers of representations per channel, learning is attenuated in these channels relative to the fastest valid channels.

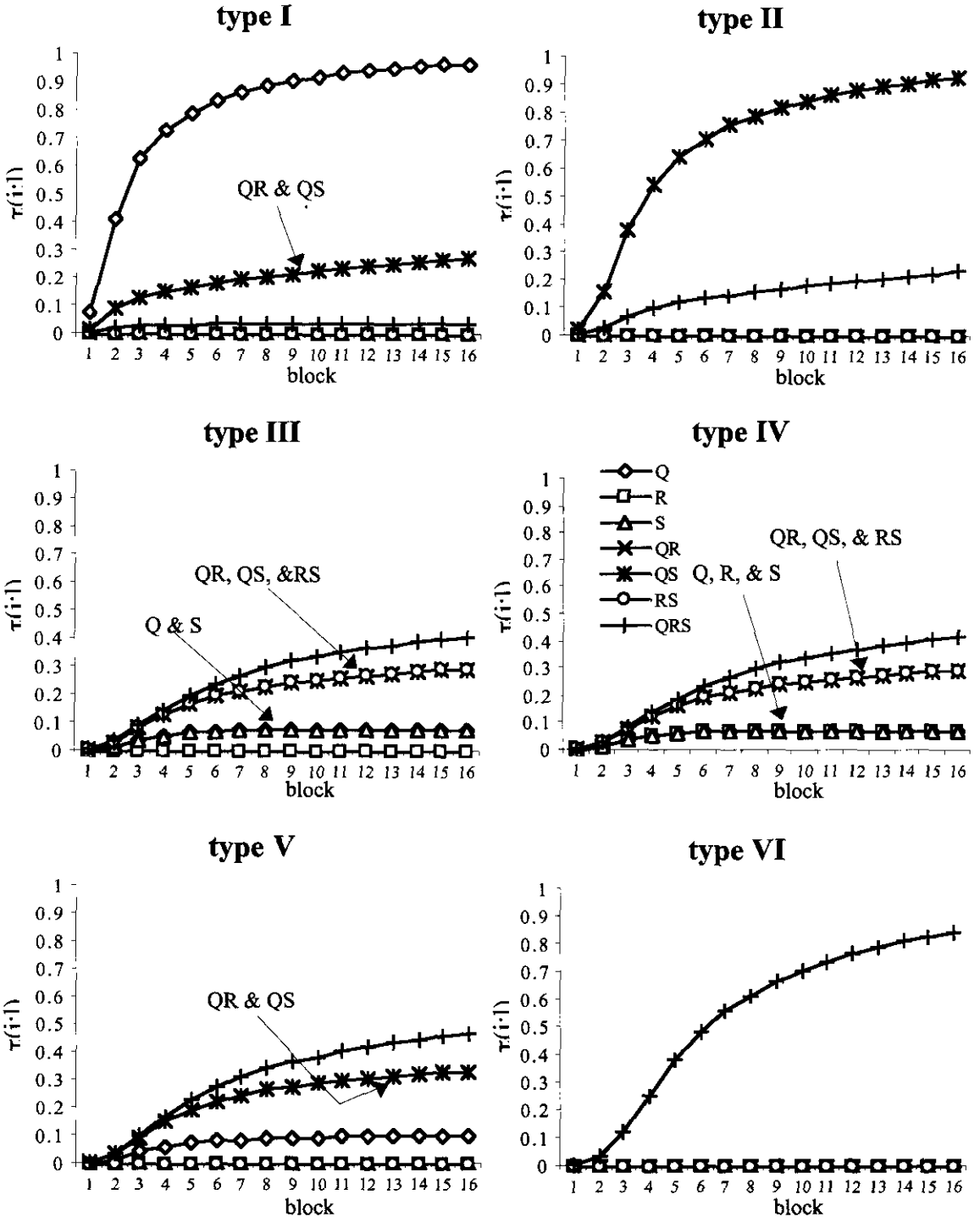

Figure 4.11: Individual channel transmission rates from the channel associability weights organisation across 16 blocks of 16 'average' trials for each of the six category structures from Shepard *et al.* (1961). Key shown for type IV applies to all graphs.

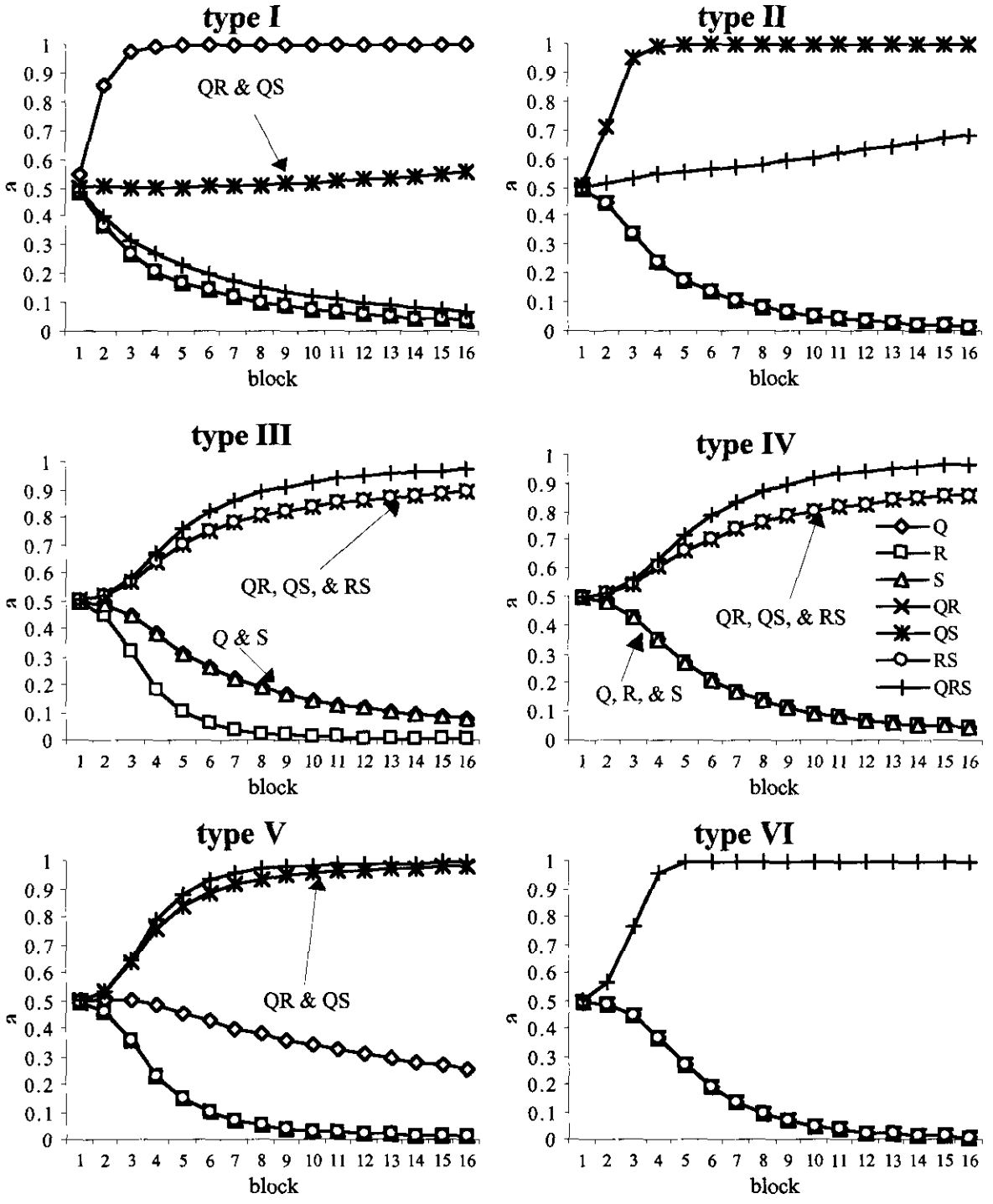

Figure 4.12: Individual channel associabilities from the channel associability weights organisation across 16 blocks of 16 'average' trials for each of the six category structures from Shepard *et al.* (1961). Key shown for type IV applies to all graphs.

4.3.3. Dimensional attention

The final approach discussed here is to locate attention parameters on the dimensions themselves. As mentioned above this leads to some conceptual problems for the configural-cue type of representation because one is obliged to suggest some way in which the dimensions themselves interact with individual representations. Consequently this type of model is the most complex of the three.

4.3.3.1 Role of dimensional weights in detector activation

One problem with the basic configural-cue model is the fact that its various representations of the stimulus are all assumed to occur with equal 'strength' on any given trial, regardless of their dimensionality. This gives the model a curious and somewhat implausible property. This is illustrated well by an XOR problem presented in the context of a three-dimensional stimulus set as in, for example, Shepard *et al*'s (1961) type II structure. This task should be easier to learn than the same problem presented with no irrelevant dimensions. This will be due to the influence of the extra redundant dimensions and the fully valid exemplar nodes that result.

Proposing some form of weights applying to particular dimensions or substitutive component pairs, requires one to propose alternative activation functions for the model such that the potentially different strengths of weight on each dimension may exert an effect.

The activation of a particular detector, in the configural-cue model, may also be described as the product of the activations of its various components. These activations are either one or zero, dependent on whether the component is present or absent. In this way, one might suggest weights on each dimension which have a multiplicative effect on the activation of components.

The problem with this approach is in terms of specifying the nature of the weights, i.e. their starting value, the range across which they may vary, and the method by which they may be altered during learning. If the weights are to represent some form of limited capacity or be normalised, for example, then problems are likely to result with regards to the activation of higher dimensionality detectors. If, for example, attention weights were to sum to one at all times then the initial activation function for representations would be $(1/n)^d$ where n is the number of dimension in the stimulus and d is the number of

dimensions in the representation being activated. This would lead to extremely small activations for higher dimensionality detectors.

The model adopted here is one that may be described as emerging from Shepard *et al*'s (1961) concept of measuring task difficulty in terms of the distribution of information across the optimum sequence, described in section 4.1.3. The 'solution' adopted is to propose that a particular representation is instantiated or activated as a result of the 'sampling' of its dimension or dimensional components. A sequential, stochastic process of sampling is suggested such that a multidimensional representation is only instantiated if its component dimensions are sampled consecutively. Figure 4.13 shows the sequential pattern of detector activation for six specific detectors, given a particular sequence of sampled dimensions.

This assumes a form of 'memory' for each representation in that, for example, a three-dimensional representation will only activate if the currently sampled dimension and the two preceding it are all different. The memory 'size' for each representation is controlled by its dimensionality, such that a one-dimensional representation is only activated if the currently sampled dimension is *its* dimension, and a two-dimensional representation is instantiated if the current and last samples are its two components. The role of attention weights in this model is to control the probability that a particular dimension will be sampled.

These probabilities are normalised such that the probability that a detector will be activated on a given trial, where all sampling probabilities are equal, will now be a function of $d!(1/n)^d$. This enables a sizeable improvement on the activation levels for higher dimensional detectors over simple product function but retains a certain amount of the simplicity. The model developed from this approach calculates the probabilities of each channel being active on a given trial using the initial sampling probabilities for that trial. The sampling probabilities are updated before the next trial using a back-propagation of error algorithm.

What this model of representation activation means with respect to the time course of processing and architectures capable of implementing it is open to question. These issues will be discussed in more detail in chapters 6 and 7, when connectionist implementations of the approach are described.

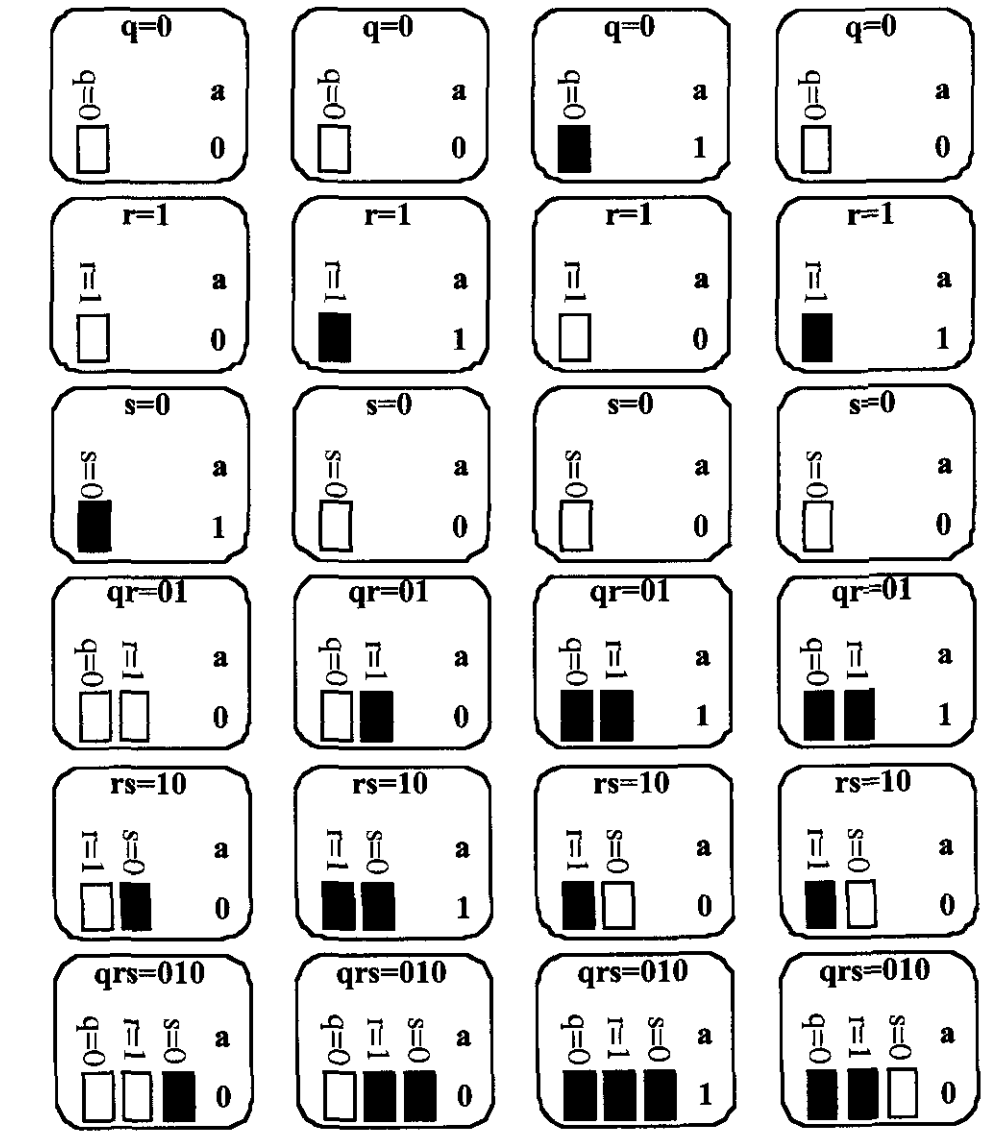

Figure 4.13: Sequential pattern of detector activation (for six detectors) for the model described in the text. Each 'component' represented by the detector (shown by labelled, filled or open rectangles) has a 'memory', in time steps, of its dimensionality (or number of components) minus one. For example, each component of the qrs=010 node will remain active for two time steps after it has been activated by the relevant input signal.

4.3.3.2 Adjusting sampling probabilities using a back-propagation scheme

This model assumes the activation scheme for representations or channels described above and illustrated in figure 4.13. In this approach, the sampling process within each trial is representable in terms of a Markov process. The probability of a particular dimension being sampled at time t is only a function of the sampling probabilities for each dimension and the dimension sampled at time t-1.

The sampling process is dependent on the vector of initial, modifiable sampling probabilities, **p**=(p(q), p(r), p(s)) and the matrix of transition probabilities **P**={p(j| i)}. The probability p(j| i) represents the conditional probability $p(X_{t+1}=j|\ X_t=i)$ which is the probability that the dimension sampled, X, at t+1 will be j given that the dimension sampled at t is i.

An additional factor affecting the transition probabilities is introduced for this model. It is assumed that the more uncertain the system is regarding which dimension to sample next (the more equal the sampling probabilities are), the less likely it is to sample to the same dimension twice in a row. This factor is introduced to enhance the probability of activation for configural channels. In a way, equal sampling probabilities are indicative of a 'requirement' for configural representations to be activated. In this model, sampling probabilities are normalised after being affected by back-propagated signals. If they remain equal after this update, the indication is that all of the dimensions of the task are equally valid.

For the Shepard *et al.* (1961) tasks, this tends to mean that configural representations are required for the task and that individual dimensions are, at best, not completely valid. Obviously, there may be other learning situations where this is not the case. For example, the dimensions may all be fully valid and, consequently, redundant. The role of this parameter will be discussed in more detail below and, more fully in the context of the connectionist models developed using this approach.

The measure used to affect the probability of consecutive sampling of the same dimension is just the entropy of the sampling 'decision', denoted H(D), it is evaluated by the following function,

$$H(D) = -\sum_{\substack{d=1 \\ d\in \mathbf{p}}}^{N} p(d)\log_N p(d) \qquad (4.23)$$

where N is the number of dimensions, d, in **p**, three in this case.

The probabilities of channel activation are determined by first determining the probabilities that each dimension is sampled in a trial. Because the probability of sampling the same dimension consecutively is partially determined by a value related to the entropy of the sampling decision, these probabilities are not the same as the initial sampling probabilities given in vector **p**.

One can derive an estimate of the sampling probabilities by multiplying **p** by $\mathbf{P}^Z$ or the transition matrix raised to a power Z. The resulting vector, $\mathbf{p'}= \mathbf{pP}^Z$ gives estimates of the probabilities of each dimension, (p'(q), p'(r), p'(c)), after Z average samples. For these models Z=7 was used.

The probability of a single dimensional channel, e.g. Q, being activated is thus just p'(q). The probability of a two-dimensional channel, c, with two dimensions i and j, being active is given by the following;

$$p(c)=\left(p'(i)p(j_{t+1}\,|\,i_t)\right)+\left(p'(j)p(i_{t+1}\,|\,j_t)\right) \qquad (4.24).$$

This is simply the sum of the probabilities of sequences ij and ji. The conditional probabilities are the entries in the transition matrix **P**. Assuming three dimensions i, j, and k the entry in the matrix corresponding to $p(j_{t+1}|\, i_t)$ is calculated as follows;

$$p(j_{t+1}\,|\,i_t)=\frac{p(j)}{p(j)+\left(p(i)(1-H(D))\right)+p(k)} \qquad (4.25).$$

The method used in equation 4.25 can be expanded to calculate the probability of the three-dimensional channel, QRS, being active during a trial.

$$\begin{aligned} p(QRS)=&\left(p'(q)p(r_{t+1}\,|\,q_t)p(s_{t+1}\,|\,r_t)\right)+\left(p'(q)p(s_{t+1}\,|\,q_t)p(r_{t+1}\,|\,s_t)\right)\\ &+\left(p'(r)p(q_{t+1}\,|\,r_t)p(s_{t+1}\,|\,q_t)\right)+\left(p'(r)p(s_{t+1}\,|\,r_t)p(q_{t+1}\,|\,s_t)\right) \qquad (4.26)\\ &+\left(p'(s)p(q_{t+1}\,|\,s_t)p(r_{t+1}\,|\,q_t)\right)+\left(p'(s)p(r_{t+1}\,|\,s_t)p(q_{t+1}\,|\,r_t)\right) \end{aligned}$$

Which is to say, the sum of the probabilities of all three-step sequences where a different dimension is sampled on each sequence, i.e. p(qrs), p(qsr), p(rqs), p(rsq), p(sqr), and p(srq).

To check that the value of Z=7 provided a reasonably accurate estimate of the average sampling probabilities, p'(d), a model of the sequential sampling process was set up using the transition matrices as calculated using equations 4.23, 4.24 and 4.25. The

model randomly determined, on each step, the next dimension sampled by comparing a random value in the unit range with the conditional probabilities for each dimension, given the current dimension.

Each model ran for 2000 iterations with a given **p** vector, determined by normalising three random numbers in the unit range. This was repeated for 1000 different **p** vectors with the relative frequency of each dimension determined for each run. These averages were compared to the values determined by $\mathbf{pP}^7$. The average absolute difference between the observed frequencies and the estimated ($\mathbf{pP}^7$) values was 0.0051 suggesting that Z=7 would provide adequate estimates for this model. The absolute difference did not appear to vary much with the value of p(d). The average absolute difference for each quarter of the range of values for p(d) was as follows: p<=0.25, error 0.0049, n=736; 0.25<p<=0.5, error 0.0051, n= 1995; 0.5<p<=0.75, error 0.0056, n=247; 0.75<p<=1, error 0.056, n=22.

The representation of sampling in terms of a Markov process does involve making certain assumptions with regards to the nature of the sampling process. The implication is that the probability of a dimension being sampled is, approximately, equal to the probability of its sampling given an infinitely long sampling period. Obviously this cannot be the case for any system having to operate in time. Consequently, the probabilities calculated represent an approximation ignoring, to some extent, the role of variance given shorter duration sample sequences. The importance of this increased variation in shorter sequences will be discussed in more detail in chapter 6.

The ongoing transmission rate of each channel is calculated in the same way as for the previous models but in this case, this rate is multiplied by the channel activation probability, prior to summation in the error function.

$$p(E) = 0.5e^{-g\sum_{c}\tau(c;l)p(c)} \qquad (4.27)$$

Note that the channel is now indexed using c. The channel activation probability is also used in determining the update to the channel's transmission rate. The function defining this increment is as follows,

$$\Delta p(correct)_c = \big(T(q,r,s;l) - \tau(total;l)\big)\big(T(c;l) - \tau(c;l)\big)p(c)\frac{1}{n_c}k \qquad (4.28).$$

The rule is similar to that used for the relative associability weights model except that p(c) is used in place of a modifiable weight as in equation 4.20.

Modifying the dimensional sampling probabilities that, in turn, control the channel activation probabilities is done using a back-propagation scheme.

$$\Delta p(d) = \left(T(q,r,s;l) - \tau(total;l)\right) k_p \sum_{d \in c} \tau(c,l) p(c) \qquad (4.29)$$

The increment to the sampling probability is, basically, the overall discrepancy multiplied by the output from each channel for which the dimension d is an element. All of this is multiplied by a rate constant k_p. These increments are added to the probabilities on that trial with the probabilities on the next trial being the normalised, incremented values from the previous trial.

4.3.3.2.1 Results

Figure 4.14 shows the increase in 1-p(E) by the model across 16 blocks of 16 trials using k=0.1, g=4, and k_p =1. The sampling probabilities were initialised at 1/3. Again, the order of difficulty reported by Shepard *et al.* (1961) is displayed by the model. The model represents learning of the type II tasks as being noticeably easier than that for the types III to V tasks. There is a slight sigmoidal shape to the graphs for all tasks, excepting the type I task. Increasing the weight learning rate k can reduce this trend but the result of this is to improve performance on all tasks. While this does not affect the overall order of difficulty represented in figure 4.14, it does narrow the gap between the type II and the types III to V tasks.

The 'selective attention' used in this model is responsible for the correct ordering in this case. Figure 4.15 shows the results for the model when the attention learning rate, k_p is set to zero with all other parameters the same as for figure 4.14. In this case performance is fairly equivalent to the basic model presented in section 4.2.2.

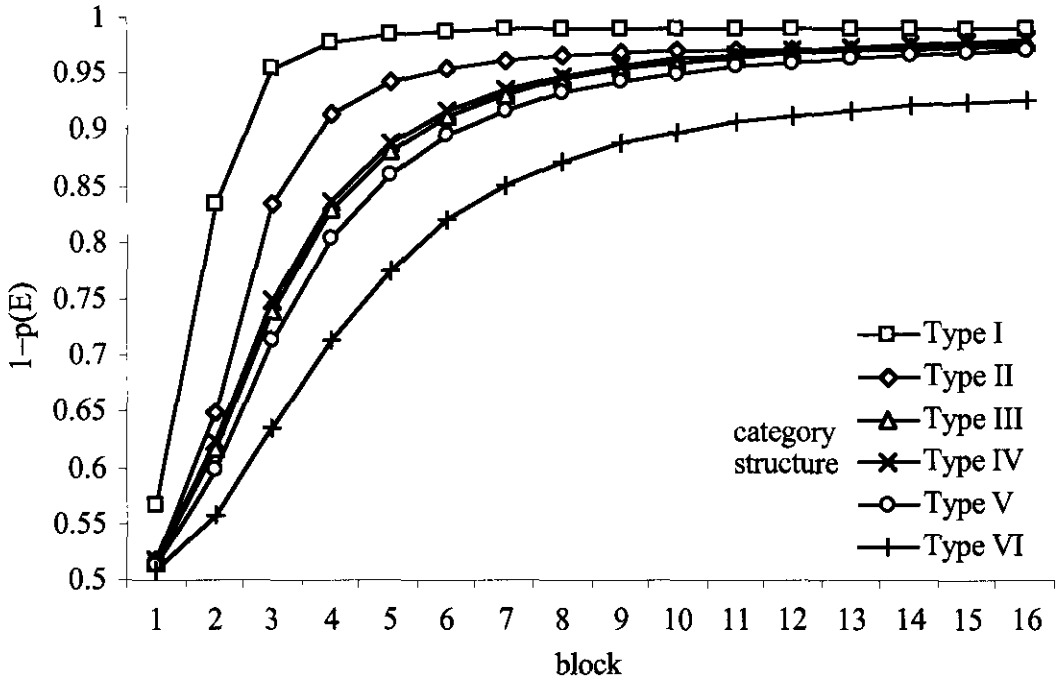

Figure 4.14: Increase in 1-p(E) by the sequential sampling configural-cue organisation with modifiable dimensional sampling probabilities across 16 blocks of 16 'average' trials for each of the six category structures from Shepard *et al.* (1961).

Figures 4.16 to 4.18 show the channel activation probabilities, transmission rates and sampling probabilities for the model with modifiable sampling rates. A number of interesting features are indicated.

The sampling probabilities shown in figure 4.18 display a marked tendency to restrict sampling to the most valid dimension(s) particularly in the type I and II structures. While the previous two models indicate transmission rate accumulating for redundant valid channels, this model shows a reduction in this tendency, illustrated in figure 4.17. Almost all of this redundant rate is accumulated during the first couple of blocks. When the frequencies with which channels are activated are taken into account, shown in figure 4.16, by about the third block of average training the contribution made by these channels will be negligible.

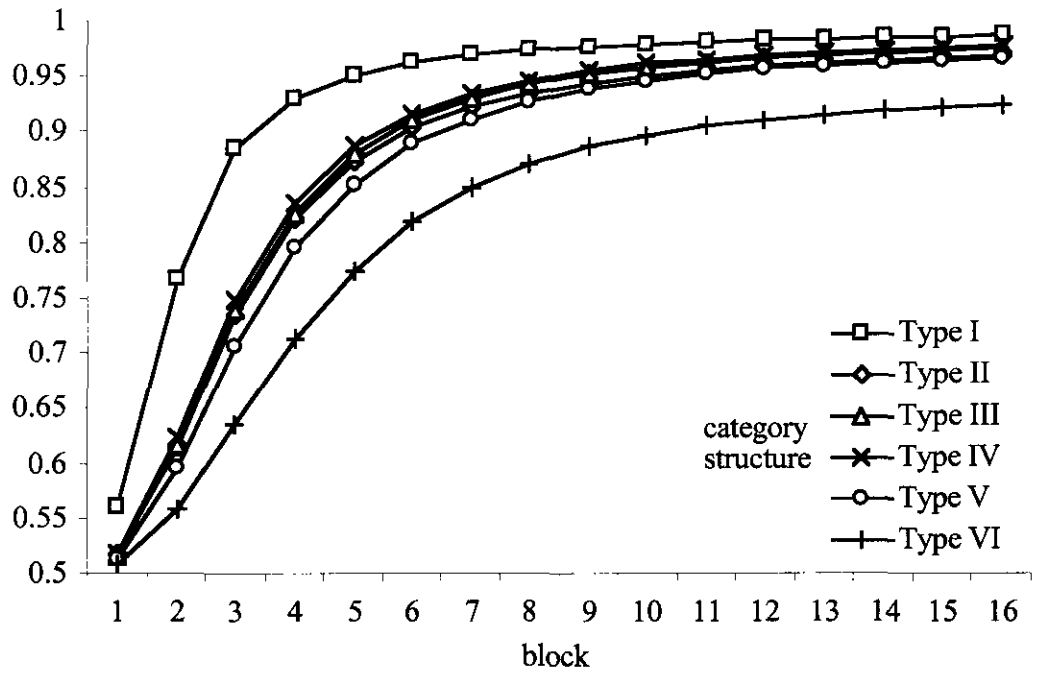

Figure 4.15: Increase in 1-p(E) by sequential sampling configural-cue organisation with learning rate k_p set to zero across 16 blocks of 16 'average' trials for each of the six category structures from Shepard *et al.* (1961).

As discussed in the previous chapter, experimental data concerning dimensional relevance shifts and blocking of conditioning suggest that dimensions that are irrelevant to the task appear to be 'blocked' with respect to subsequent learning tasks. For the first two models in this chapter some control over responding will accumulate to these dimensions via their involvement in higher dimensionality, valid, but redundant, channels. This seems to be less likely for the dimensional attention model.

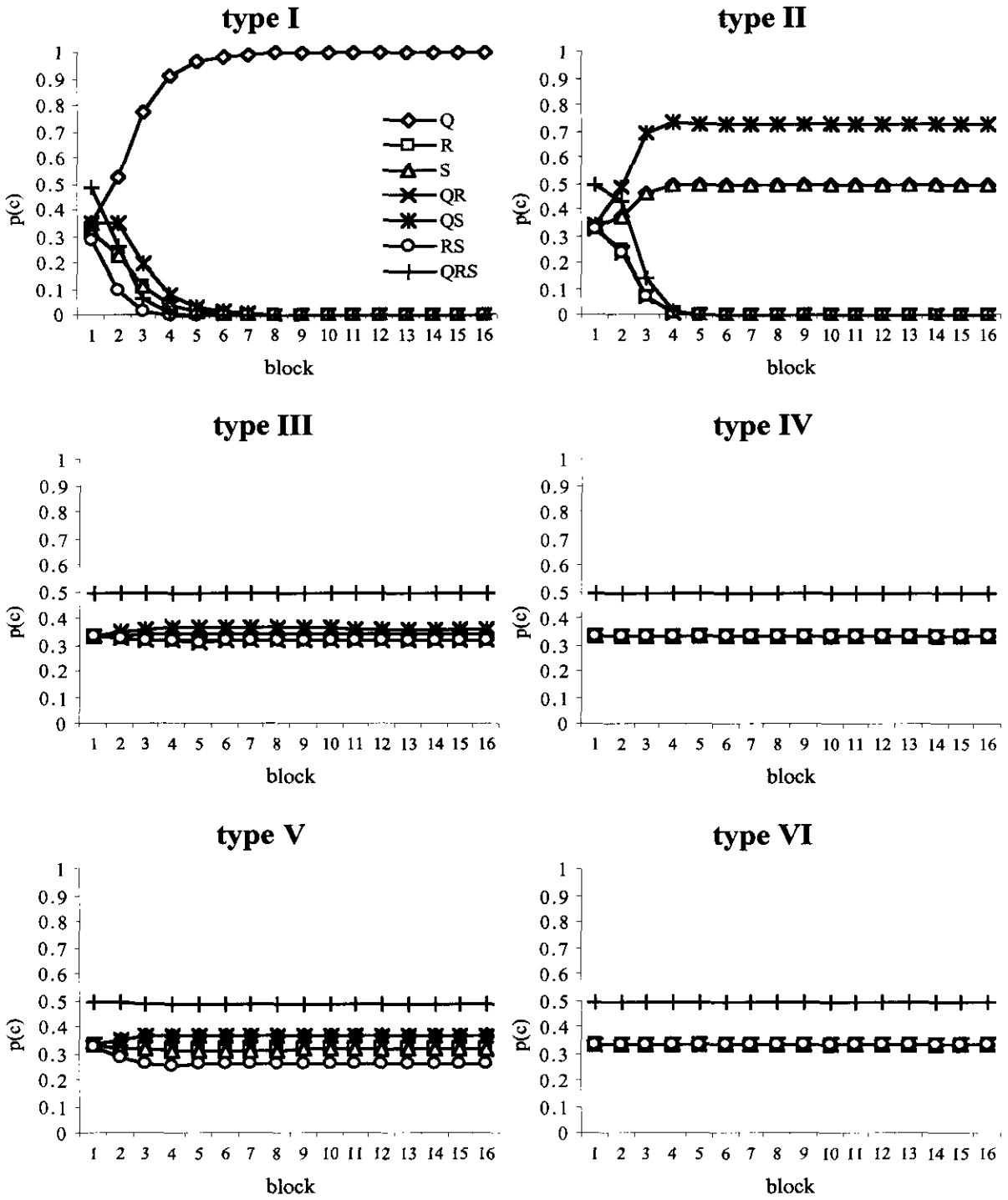

Figure 4.16: Individual channel activation probabilities p(c) from the dimensional attention model across 16 blocks of 16 'average' trials for each of the six category structures from Shepard *et al.* (1961). Key shown for type I applies to all graphs.

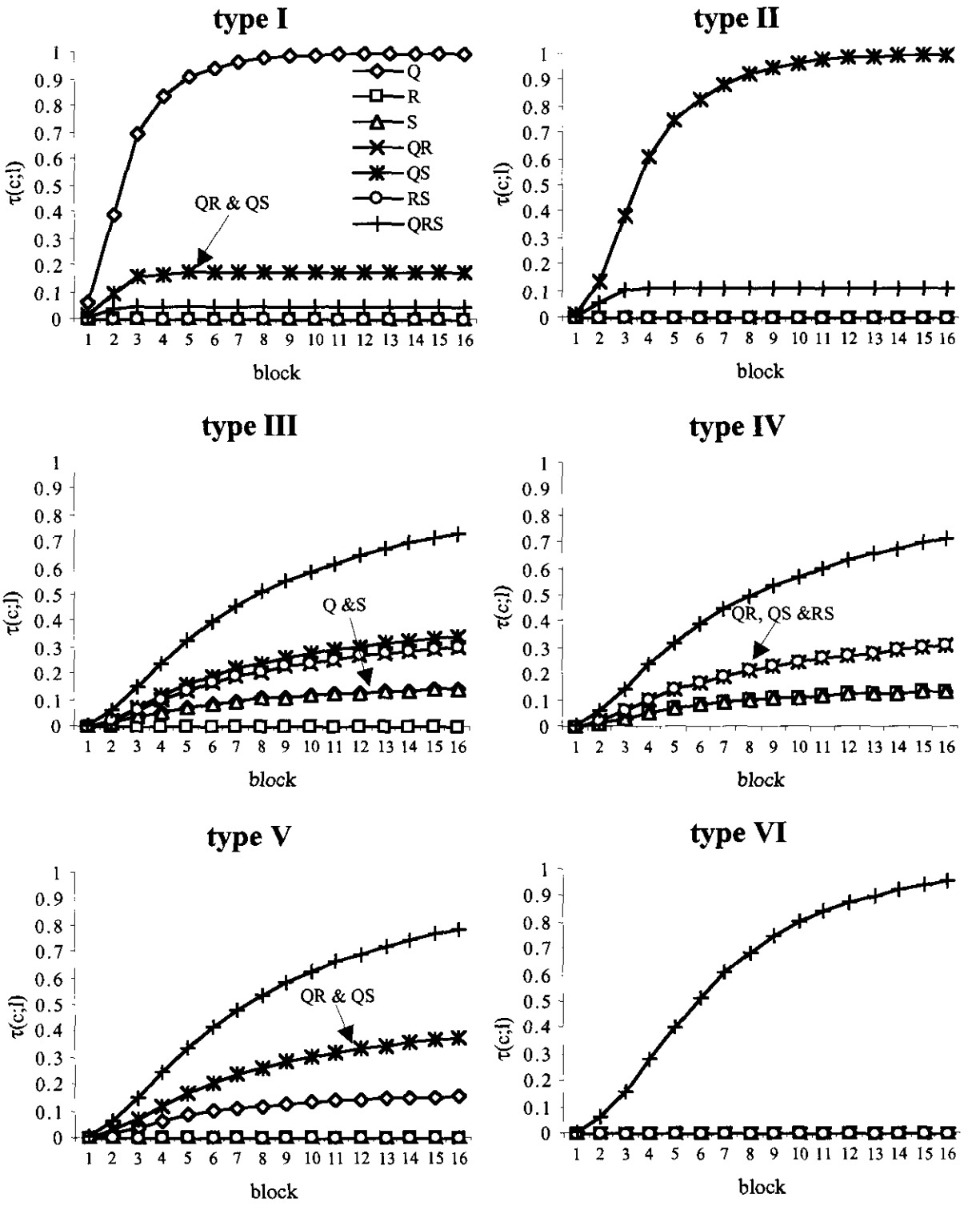

Figure 4.17: Individual channel transmission rates from the dimensional attention model across 16 blocks of 16 'average' trials for each of the six category structures from Shepard *et al.* (1961). Key shown for type I applies to all graphs.

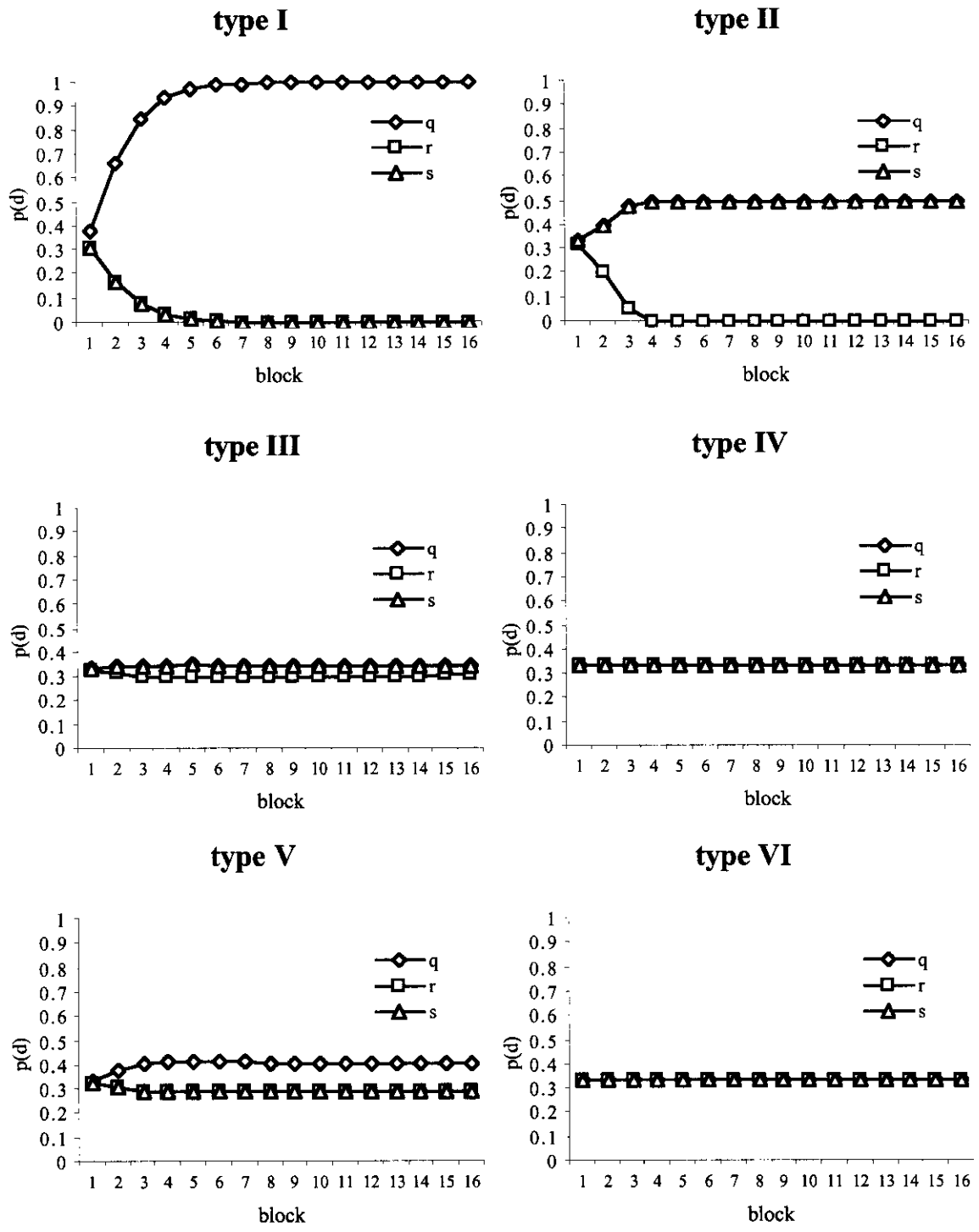

Figure 4.18: Individual sampling probabilities p(d) from the dimensional attention model across 16 blocks of 16 'average' trials for each of the six category structures from Shepard *et al.* (1961).

4.4. General discussion concerning the transmission rate approach

The transmission rate models described above are somewhat related to the stochastic learning models described in the previous chapter. Like these models they are limited with regards to the scope of their application. These models characterise the stimuli in terms of their average properties across the course of the experiment. They are consequently unable to model situations in which these properties may change during the experiment. They are thus incapable of offering any representations of concurrent learning tasks, or those which involve transfer stimuli. The models seem best suited to experiments such as the Shepard *et al.* (1961) learning tasks where the structure of the experiment is fixed.

Unlike the stochastic models described in the previous chapter, this approach does specifically relate to methods of representing the way in which stimuli may be encoded, and the relationships between various encodings and learning. The transmission rate models, as discussed above, appear to be somewhat closer to connectionist models in this respect. Connectionist models *are* effectively communication channels and as such, the form of analysis described above would appear to be a way of examining the rates at which they may learn tasks, dependent on the encoding methods they implement.

The next chapters detail attempts to implement the designs suggested here using a connectionist approach. This chapter has certainly indicated that a network using the configural-cue representation may be readily adapted to allow modelling of the data reported by Shepard *et al.* (1961) and Nosofsky *et al.* (1994). With respect to other tasks, as will be discussed, the dimensional attention model may be the model that would gain most from this approach.

A second issue, which may be addressed for all of the above models, concerns the differences between specific exemplars. In the above models specific exemplars are not examined, and so differences in their logical status in tasks III to V cannot really be investigated. It seems likely that logical status may determine the redundancy of the network with respect to individual stimuli. For example, central category members are characterised by having more valid cues and cue configurations associated with them than

peripheral or exception members. Redundancy, in this case, may well lead to different patterns of 'growth' in different spatial sub-networks dependent on whether the stimuli they are valid for are central or peripheral. As will be discussed in the next chapter, this leads to particular problems in relation to using overall error signals in the determination of weight updates, as might be required by an implementation using Rescorla & Wagner's (1972) learning rule.

Chapter 5: Modelling of the Shepard, Hovland, and Jenkins (1961) experiment using modular configural-cue networks

The previous chapter discussed the potential applicability of three different schemes for modelling category learning using a configural-cue representation. Possible solutions to the problems with the configural-cue model (Gluck & Bower 1988b, Nosofsky *et al.* 1994) with respect to this task seem to involve the addition of another set of weights to the basic organisation. This chapter concentrates on the two modular approaches developed in chapter 4.

As discussed in chapter 4, the transmission rate analysis, while useful for indicating the kinds of information that may be relevant to model performance, is limited by a number of factors. These relate to its lack of representation of the differences in the logical statuses of individual stimuli. As will be seen, these differences turn out to be important for connectionist implementations of the approaches outlined in the previous chapter.

5.1. The Independent Modular Associability Weights (IMAW) model.

The simplest model discussed in the previous chapter made use of the average transmission rates of the spatial sub-channels in a fairly straightforward way. If the output transmission rate from each channel was squared prior to summation for overall output then a qualitative fit to the observed data could be achieved. It was suggested that one way to implement this would be by locating a weight between the spatial sub-channel and the output which, in some way, tracked the average, ongoing transmission rate of the sub-network.

Figure 5.1 illustrates a part of this architecture. Similar to the DALR version of the configural-cue model, tested by Nosofsky *et al.* (1994), the 26 nodes required for the representation of the objects are divided into seven 'spatial' modules. This model is perhaps better related to 'mixture of experts' architectures (Jacobs, Jordan, Nowlan, & Hinton, 1991, Jacobs, 1997, Erickson & Kruschke, 1998) than to the DALR model, in that

both learning rate within, *and* output from, the module is controlled by modular weights. Figure 5.1 represents these weights as being located within nodes (the c or channel nodes), one for each output destination (or label), per module.

Output from the sources in the module, in the direction of the label, may be regarded as summed within the c nodes, although in practice only one of the module's sources is active at any given time. The c nodes may also be regarded as 'doubling' as local representations of the target or label node activation. As with the mixture of experts architectures, learning at the output weights of individual source nodes is *local* to the module. Each module is learning the category assignments independently of other modules. The role and nature of this local learning scheme will be discussed in more detail below.

The main difference between the mixture of experts (ME) framework and this model is that the weights responsible for gating the output of each module are also independent of one another. Contribution from each module in ME networks is generally normalised, with the weights adjusted according to the *relative* efficacy of the module at predicting the output. The ME scheme has more in common with the second model in this chapter and it will be discussed in more detail in that context.

The purpose behind this model is to implement, in the simplest fashion, the 'transmission rate squared' approach outlined in the previous chapter. The model, with just two free parameters, represents a very basic system but, as will be seen, its limitations are numerous.

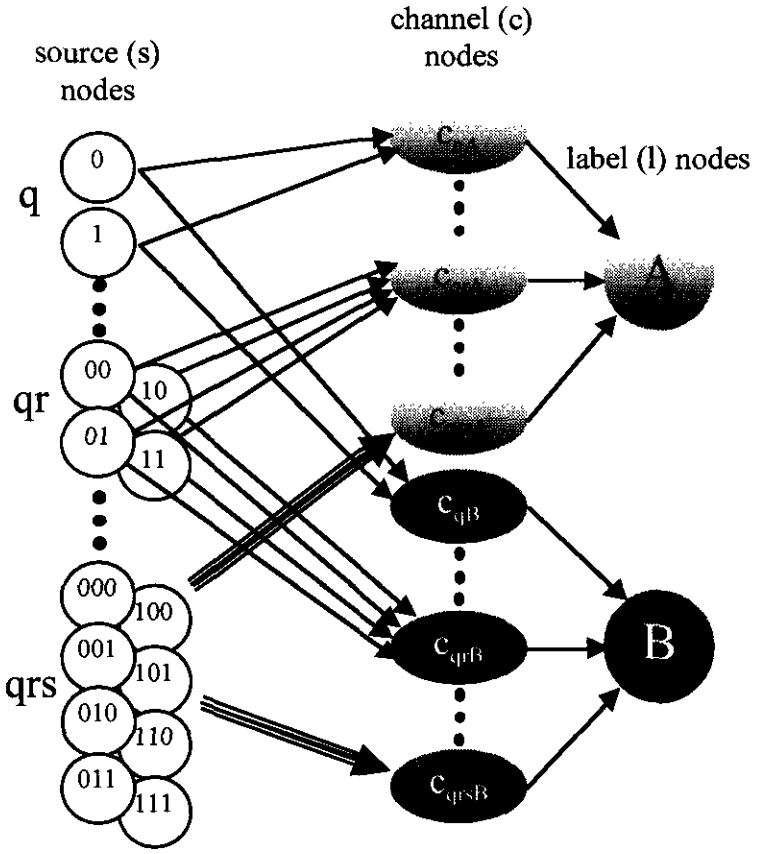

Figure 5.1: Modular configural-cue network with module or channel nodes located between source nodes and label nodes. Only three of the seven modules are illustrated for the sake of clarity. All of the qrs source nodes are connected to their channel nodes.

5.1.1. Functions defining the model

5.1.1.1 Feedforward of activation

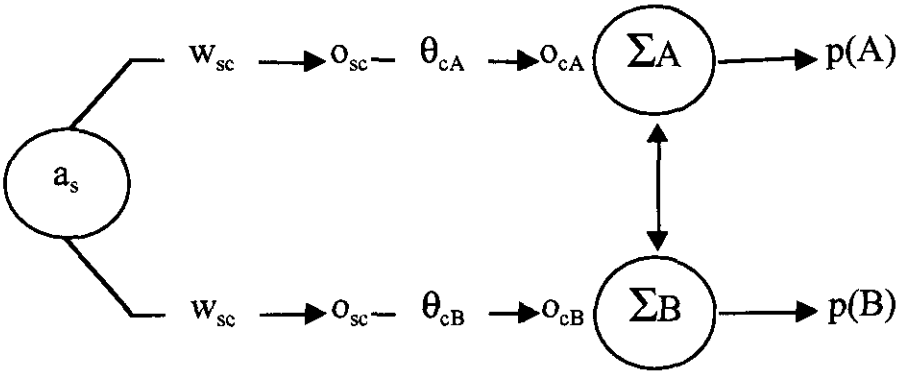

Figure 5.2: Diagram showing the feedforward pathway of activation, and notation used in the equations below from the activation of a source node s to choice probability p(*l*) (shown specifically as p(A) and p(B).

Figure 5.2 shows an example of a channel between a source node and a destination, represented by a probability of choosing one or the other category label. Weights θ_{cA} and θ_{cB} are located within channel or c nodes as shown in figure 5.1. Working backwards from the destination to the input source; the choice activation is converted to a representation of the choice probability for each label, *l*, using the normalised exponentiated output from the channel nodes. The function, described in section 3.1.2.2, and also known as a 'softmax' activation function is as follows;

$$p(l) = \frac{e^{\sum_c o_{cl}}}{\sum_l e^{\sum_c o_{cl}}} \qquad (5.1).$$

The output from the c nodes to their *l* nodes, o_{cl}, is the sum of the module's source node outputs, multiplied by the channel weight between the module, c, and the label node, *l*, or θ_{cl},

$$o_{cl} = \theta_{cl} \sum_s o_{sc} \qquad (5.2).$$

The output from the source node to the channel node c, o_{sc}, is the product of the activation of the source node, s, and the weight on the connection between s and c nodes, w_{sc};

$$o_{sc} = a_s w_{sc} \qquad (5.3).$$

As described above, only one source node per module will be active at any given time as a source node has activation, a_s, of one, if that particular cue or cue configuration is present on that trial, and zero if absent.

5.1.1.2 Weight update functions

The various weights are updated at the end of each trial. As described above, the weights between source nodes and their channel nodes are altered according to the discrepancy of the total source node output for a module and the label node activation. There is no 'global' component to the error signal, each weight begins at zero and is updated according to the following function;

$$\Delta w_{sc} = \left(\left(a_l \operatorname{sgn}(\theta_{cl})\right) - o_{sc}\right) a_s |\theta_{cl}| \lambda_w \qquad (5.4).$$

The label node, *l*, refers to the label node to which the particular c node is connected. Each s node, as shown in figures 5.1 and 5.2, is connected to two c nodes; with each c node

connected to a different label node. The activation of this node, a_l, is one if the label was present on the trial and minus one if it was absent. The multiplication of the label activation by the sign of θ_{cl} occurs because these weights can, in principle, be positive or negative. The term λ_w is a learning rate parameter that remains constant throughout the simulation. The value of this parameter controls the proportion of the discrepancy between an active node's output and the label activation removed during each trial. The effect of its variation on the overall performance of the model is mediated by the size of the other learning rate constant described below.

It is worth noting that, although this was not shown in the previous chapter, the gap between the type II and types III to V structures could be enhanced by multiplying the update signal for a module by that module's maximum transmission rate. It was assumed that this maximum transmission rate was something that would have to be learnt in a connectionist implementation. In this model this learning is effectively carried out by the θ weights. As such the decision was made to multiply weight updates by the value of the θ weight (as in equation 5.4). Alternative models were examined in relation to this and other design decisions and these will be discussed below.

The θ weights are somewhat different in that they must be initialised at some absolute magnitude which is greater than zero. As equation 5.4 uses the sign of θ and its absolute value, a zero initial value will preclude the possibility of any learning at all.

As discussed above, their function is to track the average transmission rate of the module. The way this is done here may be described in terms of a connection *back* between the label node and the module. The learning process is one where the label nodes are attempting to learn to predict the input to the modules' c nodes, with the θ weight being the connection strength which results from this process. In this case the weights are updated according to the discrepancy between the guess made by the label node, via the θ weight, and the 'bottom-up' input to the c node. The function describing the change in these weights is as follows;

$$\Delta\theta_{cl} = \left(\sum_s o_{sc} - \tanh(\theta_{cl} a_l) \right) |\theta_{cl}| a_l \lambda_\theta \qquad (5.5).$$

The parameter λ_θ is a learning rate parameter that remains constant throughout the simulation. This parameter, like λ_w, controls the rate at which the weight changes in the direction of the target, in this case the target is the weighted output of the module's source nodes, on a given trial. The effect of varying λ_θ on the model's performance is related to the value of the other learning rate parameter. The nature of this interaction shapes, to some extent, the model's performance in general and will be discussed in more detail below.

The use of the hyperbolic tangent of the weighted label activation tends to enhance weight values of fully valid channels, where the weights between source and channel nodes head towards absolute values of one. In this case, the hyperbolic tangent will tend to guarantee that the output of the module's source nodes is higher than the 'output' of the label node to module connection.

An alternative but equivalent weight update function is given in Bartos and LeVoi (2001); its form in relation to the notation given here is,

$$\Delta\theta_{cl} = \left(\left(a_l \sum_s o_{sc} \right) - \tanh(\theta_{cl}) \right) |\theta_{cl}| \lambda_\theta \qquad (5.6).$$

Equation 5.5 is probably more satisfactory as it allows for the possibility of no learning occurring in the absence of feedback (when a_l may be zero). In addition it captures the idea of the transmission rate being squared in a more intuitive way.

The transmission equations given in the previous chapter are symmetrical, in that the maximum rate of a channel between combined module output and category label is the same as the rate of a channel between the label and the combined source node output. Implemented using equation 5.5, the 'squaring' of transmission rate may be conceptualised in terms of the feedforward transmission rate of the module (learning the label) being multiplied by the transmission rate of a feedback channel between label and module. The transmission rate of this feedback channel changes according to how well the label may be used to predict the category membership of the sources within the module.

5.1.2. The experimental simulation

The model was tested on each category structure twenty times with a different randomised order of input presentation for each experiment. The blocks were organised in the same way as the Shepard *et al.* (1961) experiment for the trial-and-error learning of the

individual category structure (experiment 1) and the Nosofsky *et al.* (1994) replication (see section 2.2). No attempt was made here to investigate the transfer of training effects, so the results reported below are just averages for the twenty simulations of one run through sixteen blocks of training data.

The parameters used for the simulations were λ_w=0.35 and λ_θ=0.2. The initial value of all θ weights was set to 0.5. These settings were not optimised to reduce difference between model and human performance. Varying the parameter settings generally did not effect the overall ordering of the task difficulties. The effect was generally to alter the convergence rate across all tasks and, consequently, to affect the extent to which curves were 'bunched-up' or separated. Also the values of the learning rate parameters will affect the stability of the model or the ability of the model to converge at all. The role of these particular parameter settings will be discussed below.

5.1.3. Simulation results

Figure 5.3 shows the average probability of correct responding by the model for the six category structures. The results compare favourably with the results of Shepard *et al.* (1961) and Nosofsky *et al.* (1994) (see figure 2.3), in terms of the number of errors likely to be made during the learning of each problem. There is a clear advantage for the type II structure over types III and IV, indicating that the model more closely simulates human data than the basic configural-cue model (Gluck & Bower, 1988b) and the variants tested by Nosofsky *et al.* (1994).

Comparison of this figure with figure 2.3, however, reveals a number of differences in terms of the fine detail of the learning curves. These differences and their implications for the approach will be discussed in detail below.

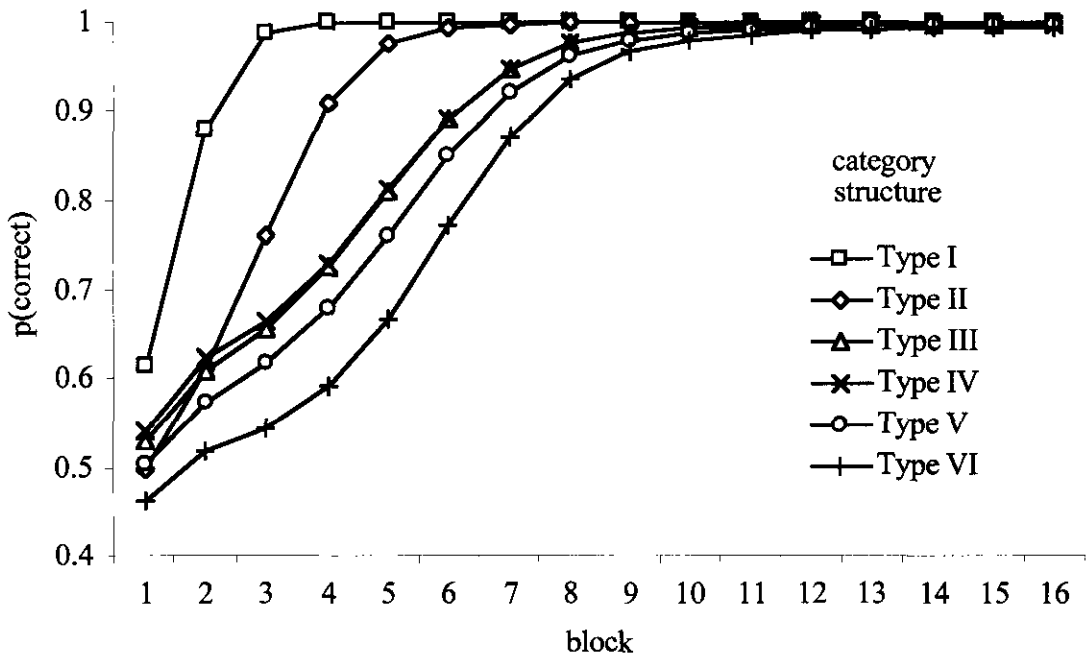

Figure 5.3: Average probability of correct output per block for the independent modular associability weights model for each of the six category structures of Shepard *et al.* (1961). Averages are based on 20 runs through the 16 blocks, on each structure with randomised order of input patterns.

As figure 5.4 shows, the model also successfully demonstrates the differences in performance on the different patterns for task types III to V. Central members of categories were learnt with lower average probabilities of error than peripheral members. The exception members in the type V structure were learnt more slowly than central and peripheral members. There were, as expected, no significant differences between pattern performance within structures I, II, and VI.

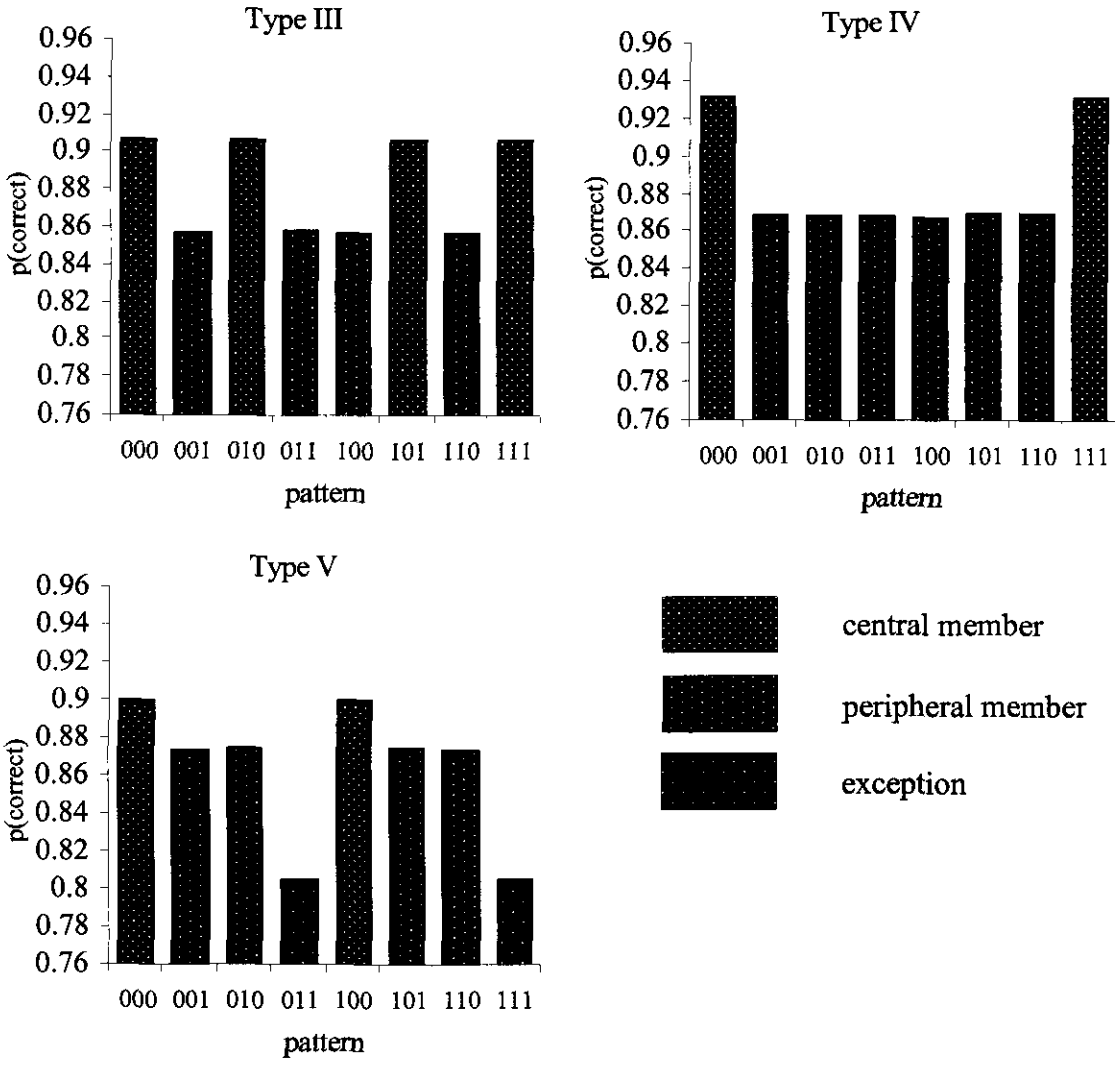

Figure 5.4: Performance of the model on individual patterns for category structures III to V. Performance is indexed in terms of the average probability of correct responding across the entire 16 blocks of the simulation. See figure 2.2 for illustrations of the relationships between these three types of member within each category structure.

5.1.4. Discussion of the IMAW model

While the above model displays a qualitative fit to the relative difficulties reported by Shepard *et al.* (1961) and Nosofsky *et al.* (1994), the way in which it produces this fit seems likely to limit its ability to address data from other learning experiments. Some of these problems generalise from the modular approach, as discussed in the previous chapter;

others pertain to something of a lack of 'robustness' of the qualitative fit under conditions where the model's parameters are varied.

5.1.4.1. Local versus global teacher signals

In common with other modular approaches (e.g. Erickson & Kruschke, 1998) this model employs a 'local' learning rule for the incrementing of weights within modules. This rule does not take into account the discrepancy of the summed output from the target value when calculating weight updates.

Different weight-update functions were considered and tested in the development of this model. These involved different uses of 'global' teacher signals; i.e. those dependent on measures of discrepancy between summed contribution to a label or *l* node and the feedback activation of that node. These variants included:

1. Global teacher signal updates for the source to channel node weights, with the θ weight update functions as in the above model.
2. Global teacher signal updates for the source to channel node weights, with the θ weight update function additionally gated by the absolute magnitude of the global teacher signal on that trial.
3. Local teacher signal updates for source to channel node weights, with the θ weight update function additionally gated by the absolute magnitude of the global teacher signal on that trial.
4. Local teacher signal updates for source to channel node weights additionally gated by absolute global error signal value, with the θ weight update function additionally gated by the absolute magnitude of the global teacher signal on that trial.
5. Local teacher signal updates for source to channel node weights additionally gated by absolute global error signal value, with just the local θ weight update function described by equation 5.5.

Note that the transmission rate version of this model, described in the previous chapter, employed a form of global teacher signal in terms of the average remaining ambiguity of the channel (maximum possible transmission rate, minus the combined transmission rate). Using global teacher signals proved problematic for this connectionist model. While the results from the above models will not be discussed in detail, the ability

of these models to represent the qualitative trends of the learning curves in figure 2.3 was somewhat worse than the model described in detail above. The orders were, generally, preserved although there was a marked tendency with some variants for performance on the type VI structure to overtake or equal performance on the types III to V structures. On other variants, this problem was made worse, or replaced by a tendency for the types II, III and IV learning curves to be very close together.

Where implemented on just the s to c association weights, the general effect of global teacher signals was to reduce the performance on the types III to V structures, relative to performance on the type VI. The reason for this is that the different logical status of the stimuli in these structures leads to a differing dependence of the network on the fully valid three-dimensional module. Because the teacher signals to this module would vary in size, dependent on the particular stimulus presented, the θ weights for the three-dimensional module would generally decrease when exposed to central members of the category (due to the lower weights in this module for these stimuli). They would generally increase when exposed to peripheral or exceptional members. The net result is an attenuated development of this module's contribution to the decision process for all stimuli.

Because the θ weights for the partially valid two-dimensional modules tend towards values of 0.5 (reflecting the average output from these modules), performance on stimuli where less than two valid two-dimensional sources are present, e.g. peripheral and exceptional members would be attenuated. These variants would thus be fairly slow at learning in the later phases of training. This problem would not be suffered for the type VI structure, where the three-dimensional module would increase its output at the same rate for all inputs.

As can be seen from figure 5.5, for the local learning model presented here, the rate at which θ weights change for modules, and the levels they reach by the end of training, depends only on the diagnosticity of the module itself. As such, for example, the θ weights in the three-dimensional module grow at the same rate and reach the same maxima on all tasks.

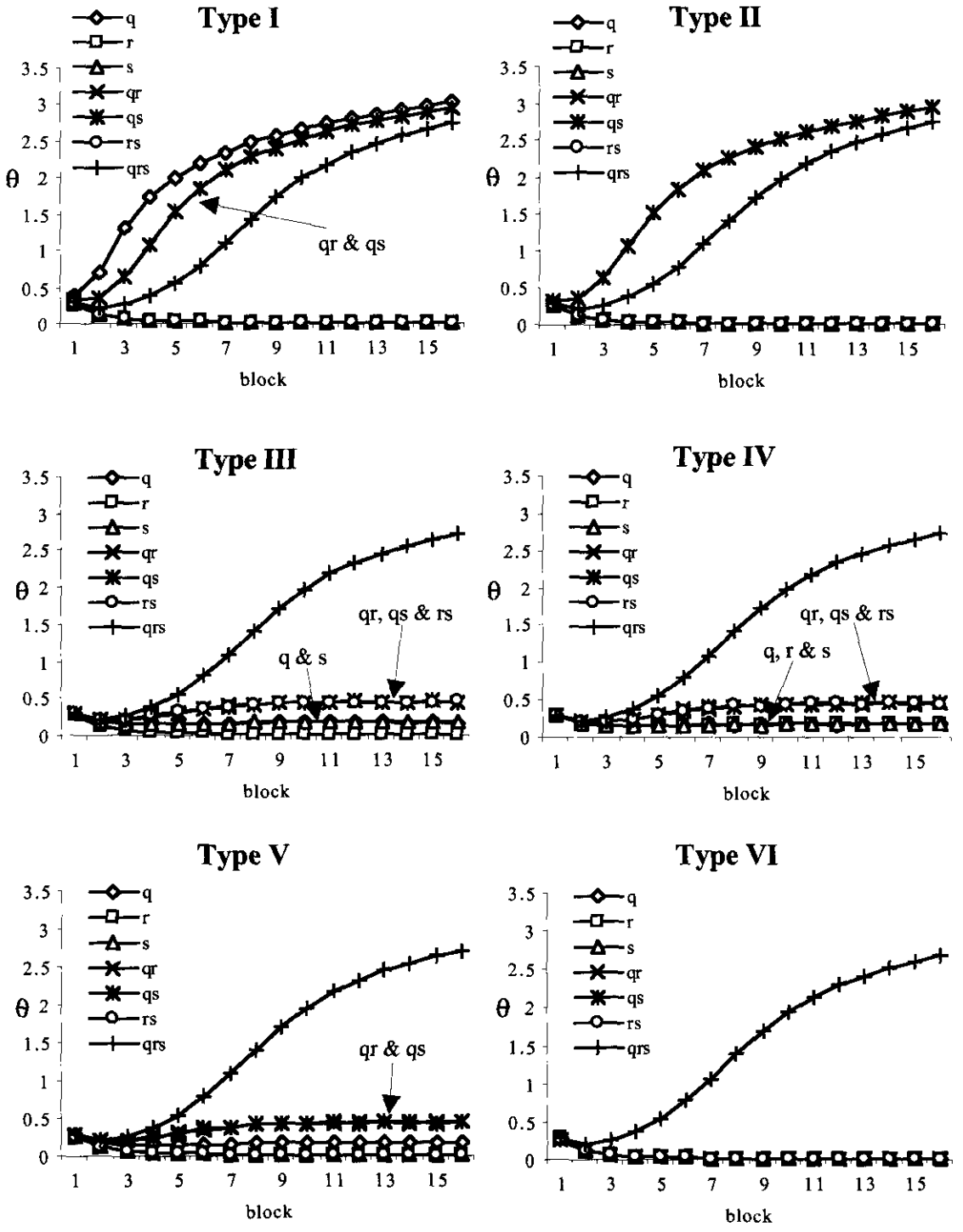

Figure 5.5: Development of θ weights in the model per block on the six category structures, averaged across the twenty simulations per structure.

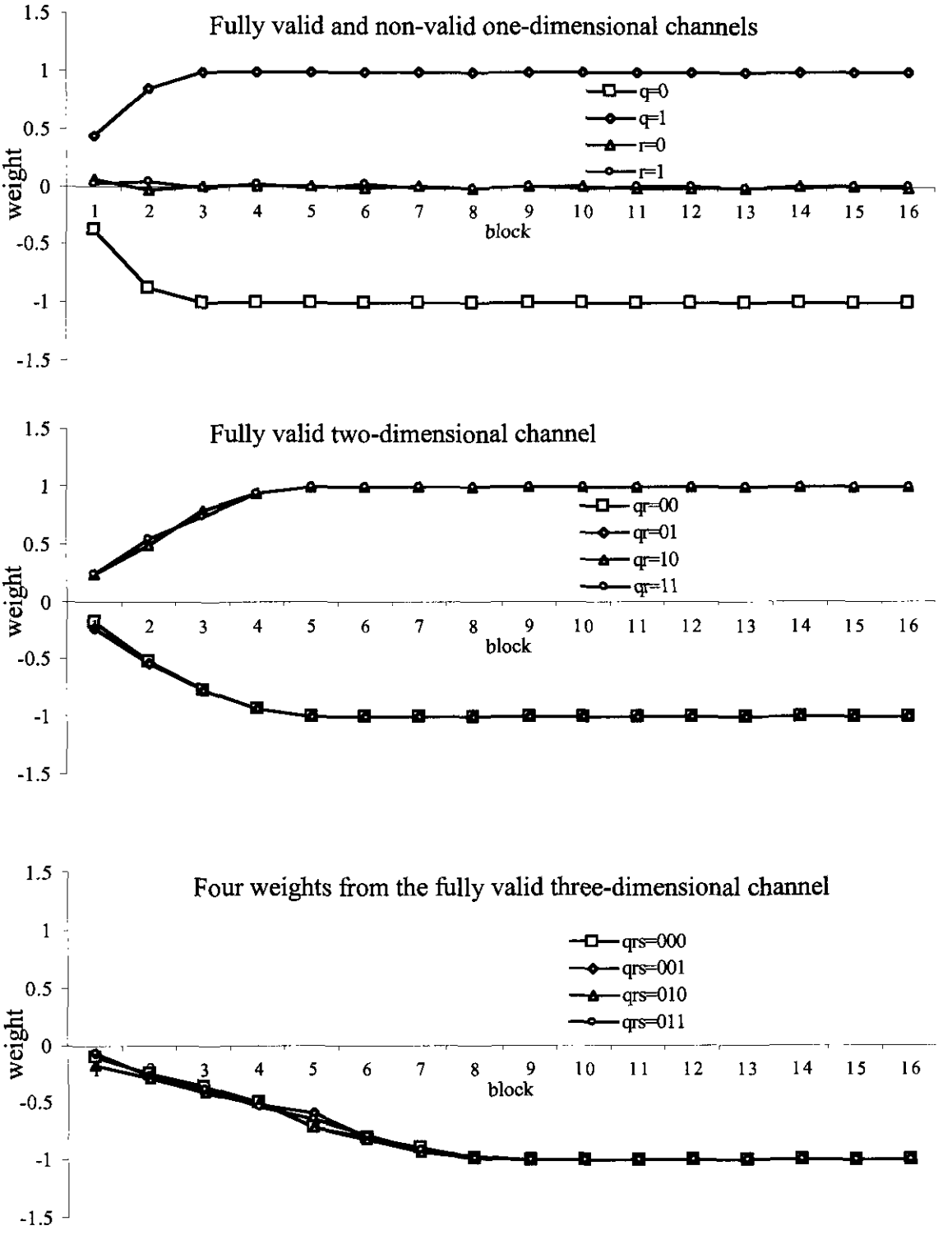

Figure 5.6: Examples of development of source node to channel node (category B) weights across the sixteen blocks of a typical run through the type I category structure.

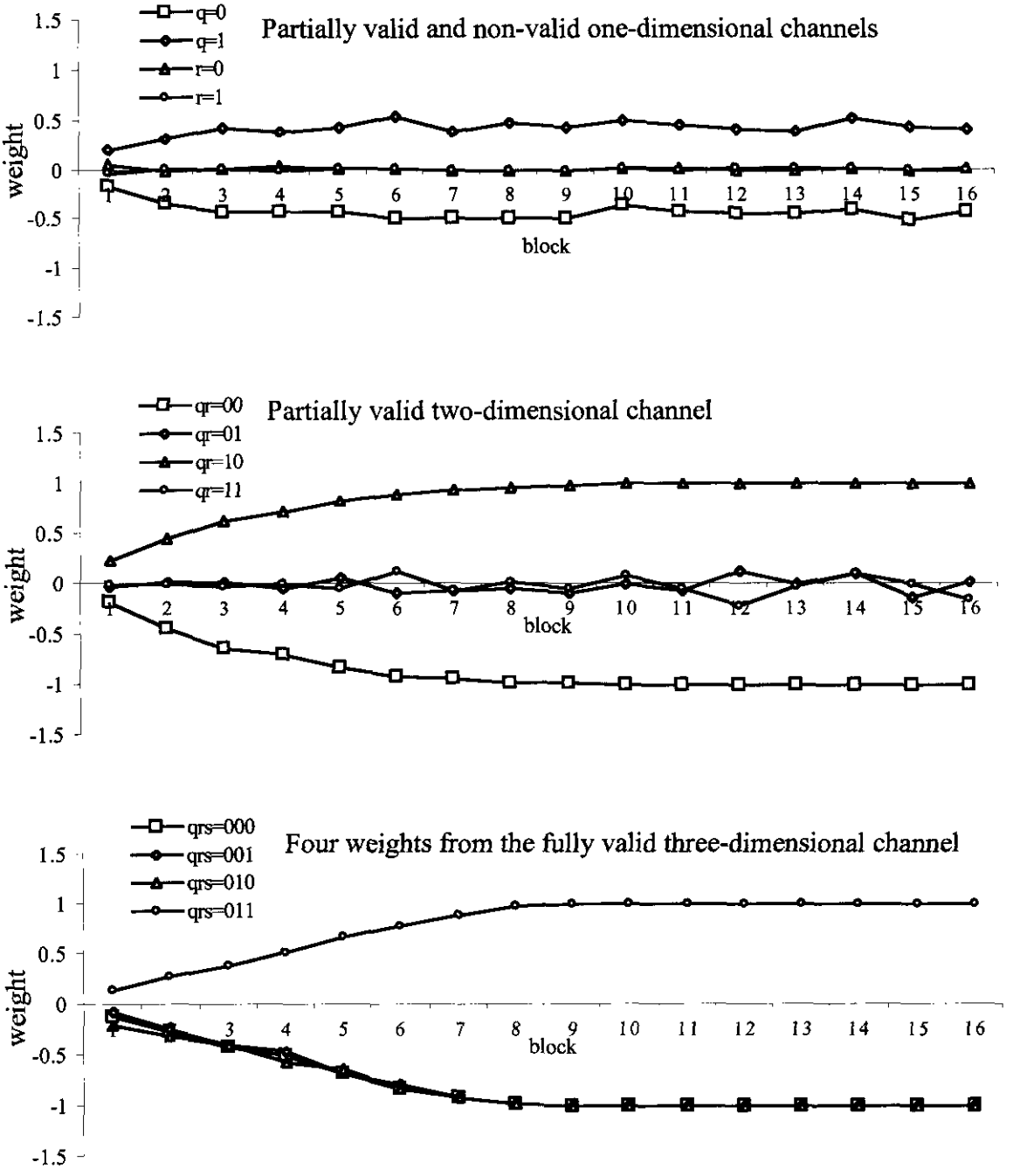

Figure 5.7: Examples of development of source node to channel node (category B) weights across the sixteen blocks of a typical run through the type V category structure.

Figures 5.6 and 5.7 show examples of the development of association weights for some of the modules in one run through the types I and V category structures. As can be seen, all of the sources in the three-dimensional module grow at the same rate towards the same maximum, regardless of the logical status of the stimulus they represent.

One might suggest that this problem might be ameliorated by the inclusion of a global rate parameter in the θ weight update algorithm. In this case the θ weights would not decrease as much in the three-dimensional module in the presence of central members simply because the global signal will be smaller, owing to the contributions of lower dimensionality modules.

This, unfortunately, results in a different difficulty. The problem observed in this case is that the difference between the type II and the types III and IV structures decreases. Much of the reason for this is attributable to an increased contribution made by the partially valid one-dimensional modules in the types III and IV structures.

As can be seen from figures 5.5 and 5.7, the weights associated with these channels in the local learning model tend to remain low. As the 'positive' teacher signals associated with stimuli for which the module is valid decrease with learning, the negative signals associated with their being invalid *increase*. Despite the lower relative frequency of the negative signals (0.25) their increase in size results in the weights stabilising, on average, at the values observed in figure 5.7. The same pattern is true for θ weights, which will decrease more when the channel is wrong than they increase when it is right.

Introducing a global teacher component to the θ weight learning rule will disrupt this pattern, as the size of teacher signals will generally decrease as learning progresses. This results in the balance between positive and negative signals for these channels being disrupted. The net result is that both associative and θ weights increase above the levels observed for the local learning model. As with standard networks deploying the Rescorla-Wagner rule, what tends to occur is increased weight gains for valid sources when their activation is accompanied by an active source that indicates the opposite decision. As such there is a tendency for the θ weights of the two-dimensional, partially valid modules (which may contain sources perfectly valid on occasions when the one-dimensional channels are 'wrong') to develop faster.

What tends to be observed with variants where only the θ weights have global learning parameters, is that on types III and IV early learning rates are much higher than that observed for the type II. Eventually, *actual* maximum transmission rates are reached by the lower dimensionality modules and this advantage disappears. The performance on the type II overtakes performance on types III and IV. While late superiority for type II is observed in the model described here, it is much more marked in these variants, with the superiority of type II generally being less than that shown in figure 5.3 following its emergence. This issue will be discussed further below in relation to the IMAW model.

In variants which combine local learning with global parameters in the associative weight update, the problem associated with purely global learning remains, with learning in types III to V attenuated relative to type VI in later blocks. Putting global parameters on both learning rules (as in variant 4) combines both problems, learning of types III and IV is more rapid than that for type II for up to three blocks, but then gets worse than learning for type VI in later blocks. In addition, asymptotic performance on all tasks is fairly poor.

This would appear to be a significant problem for this model. It implies that individual modules will learn any problem according to their individual relationship with the problem structure. Without modifications this would appear to preclude the possibility of the model being able to represent blocking of conditioning (as discussed in chapter 3).

5.1.4.2 Late superiority of type II structure

An additional problem observed for this model is the late superiority of performance on the type II structure, relative to performance on structures III and IV. As discussed above, this problem is exacerbated by the effects of introducing a global teacher signal weighting for the θ weight update function.

The principal reason for its occurrence in this model, and its more severe manifestation in the variants discussed above, is related. As described in section 5.1.1.2, the choice to include the absolute value of θ in its own learning function and the learning function for the association weights, makes it necessary to initialise the θ weights at a value greater than zero. Reducing this initial value tends to result in increasingly sigmoidal learning curves, with an overall increase in the number of blocks of training required before asymptotic performance is reached.

As also discussed in section 5.1.1.2, while the transmission rate model described in the previous chapter did not include a parameter in its learning functions which captured the channel's transmission rate, omitting it in a connectionist implementation resulted in a worse fit to the data. The resultant variant employed a local learning rule for association weights, with the response strength passed to the decision nodes weighted by the θ weights. These weights could be initialised at zero and develop using a similar function to equation 5.5, but *without* their own value included in the function.

The result was a correct ordering but, as stated above, the difference between performance on type II and types III and IV was smaller than that shown for this model in figure 5.3. In addition the difference between the type V structure and the types III and IV increased. In fact the difference between these curves was equal to or sometimes greater than the difference between types III and IV, and type II.

These factors led to the decision to multiply the weight updates, as well as the output, by the value of the module's θ weight. As discussed above this meant that the value of this weight had to be initialised at a value of greater than zero. This factor is principally responsible for the problem of the late superiority of the type II structure over that of types III and IV.

The reason for this is that, during the initial trials with the types III and IV structures, the values of the θ weights on the partially valid two-dimensional modules are actually at the level they can be expected to reach at the end of training. This means that early learning in these modules will be at an identical rate to learning in the fully valid two-dimensional module in the type II structure.

Because there are more fully valid two-dimensional sources in the types III and IV structures than in the type II, early learning may be expected to follow the same pattern as that observed for the standard configural-cue model (Gluck & Bower, 1988b). Note, in figure 5.5 early learning for all structures tends to be characterised by a drop in the value of θ weights from the initial value of 0.5. This is to be expected; the weights begin to increase again when the values of association weights in a module exceed the hyperbolic tangent of the θ weight for that module.

Figure 5.5 indicates that this seems to happen by about the third block. At this point, the superior validity of the fully valid two-dimensional module in the type II

structure enables its θ weight to increase at a higher rate than the θ weights for the modules in structures III and IV.

As suggested above, this problem is related to the problems associated with the use of global teacher measures on the θ weight learning algorithm. Both problems are a result of the θ weights not accurately representing the effective transmission rate of the module they pertain to. The more 'equal' these weights are, the more learning in the model will resemble that of the standard configural-cue model.

5.1.4.3. Overview and possible developments to the model

While the IMAW model produces what may be described as superior qualitative fits to the Shepard *et al.* (1961) and Nosofsky *et al.* (1994) data than the standard configural-cue model, it would appear to have a number of problems. These problems, to some extent, compromise its performance on this task but also would appear to preclude its generalisability to other learning tasks.

The discussion above would appear to indicate that the 'tricky' aspect regarding getting the model to show the desired order of task difficulty, is one of keeping the types III to V learning curves between the curves for the types II and VI tasks. Certain modifications would appear to close the gap between type VI and types III to V, while others seem to enhance type III and IV performance relative to type II.

In the context of the corresponding transmission rate model, presented in the previous chapter, this would appear to be a problem of making sure the combined transmission rates of the sets of partially valid modules in the type III and IV structures sum to less than one. All of the structures have a fully valid three-dimensional module. This is a task, in the case of the type IV structure, of trying to guarantee that the sum of contributions from three partially valid one-dimensional modules and three partially valid two-dimensional modules adds up to less than the contribution of a single, fully valid two-dimensional module.

The way in which this model appears to achieve this may be criticised as being ˥mewhat contrived, as it basically involves raising the transmission rates of modules to ˢ greater than one. This is obviously most likely to affect performance on the types ι ˥ctures, as the transmission rates of all but one of their modules are less than

one. Squaring, for example, will not affect the type II structure as much because the maximum transmission rates of its modules are all unity.

The model may also be criticised on the grounds that its ability to generalise to other tasks would appear to be quite limited. It will be able to demonstrate, for example, compound-component and component-compound transfer, by virtue of its configural-cue representations. It would, however, have difficulty with a compound-component discrimination problem. The presence of the component would have, on average, no validity with respect to the prediction of the outcome. The model would have to represent the absence of one component and the presence of the other as distinct two-dimensional sources in order to be able to do this task. Whether this is a psychologically valid way to represent stimuli is somewhat questionable, as it does suggest that the absence of anything may be accompanied by distinct sources which activate when that thing is absent.

Blocking of conditioning appears to be another phenomenon that this model would have problems representing. The lack of global error parameters precludes recourse to a 'lack of learning' explanation of the phenomena as modelled by the Rescorla-Wagner learning rule (Rescorla & Wagner,1972). The model was not developed as a means of addressing these observations though, and it may be possible to alter it such that some representation may be offered. The initial values of the θ weights in the model, for example, may be determined for this type of concurrent task as a function of the global error present. If all θ weights began at some greater-than-zero level, then when a module is instantiated in some way by input blocking would obviously be difficult to represent.

If, however, the initial θ weights were dependent on the error, or magnitude of the teacher signal, following the presentation of the new configuration, then learning in the configural module, and the module containing the new redundant relevant source, would be severely attenuated. If the other problems of the model can be overcome this could be an avenue for further research.

5.2. The Relative Modular Associability Weights (RMAW) model

The second approach discussed in the previous chapter was to weight a channel or module's contribution to the overall category judgement by some measure of its *relative* transmission rate. This measure is, in a general way, comparable to Mackintosh's (1975) concept of cue associability. The second model in this chapter represents an attempt to implement this idea in a modular connectionist network.

The basic idea behind Mackintosh's model is that there is an adaptive learning rate parameter associated with each cue. This parameter adapts according to how well the cue is predicting the outcome in relation to other cues present on that trial. If the cue has greater association strength than the combined strength of all of the other cues present on that trial, then its learning rate parameter increases. If its strength is equal or lower, then its associability decreases.

As discussed, however, this concept related to individual dimensions or cues. The model outlined by Mackintosh made no commitment to handling the configural or exemplar representations that seem essential to the modelling of tasks such as the Shepard *et al.* (1961) category learning tasks. This model represents the parameters as a characteristic of the module or spatial channel. The model, therefore, has a similar structure to that illustrated for the previous model in figure 5.1. The channel or c nodes, in this case, 'house' the associability weights, the evolution and nature of these weights being the principal difference between this model and the IMAW model.

As with the previous model, learning 'within' modules is local to that module. The problems discussed in the context of the previous model, with regards to global error signals, applied equally to this model when variants were tested.

The problems of the previous model with regards to the accumulation of transmission rate in redundant relevant channels would, it was hoped, be eliminated by the use of the relative associability weights as a variable in the associative weight update functions. Because the relevant channels would accumulate association strength at a rate proportional to the validity of their cues *and* the frequency of those cues, high dimensional relevant channels would generally have less associative strength than low dimensional

ones. The result would be that the associability of these high dimensionality channels would generally be negative in relation to the lower dimensionality ones.

In addition, the fact that there would be some *active* reduction of the contribution of non-valid channels (unlike the passive reduction for the previous model) may enable the superiority of the type II structure to be revealed earlier in learning.

5.2.1. Functions defining the model

5.2.1.1 Feedforward activation

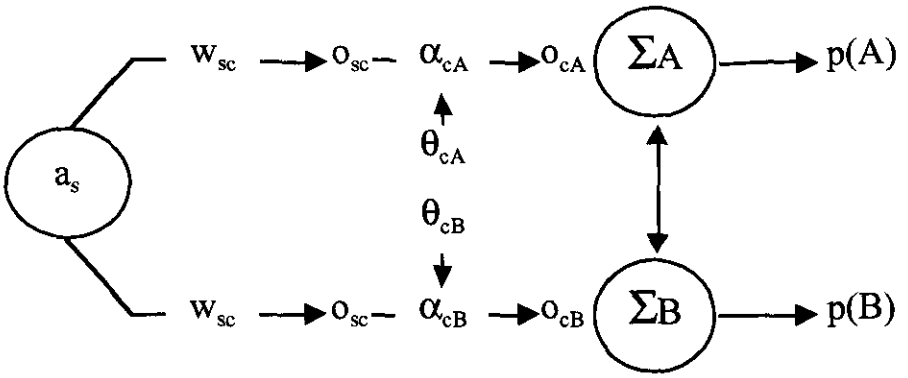

Figure 5.8: Diagram showing the feedforward pathway of activation and notation used in the equations below from the activation of a source node s to choice probability p(*l*) (shown specifically as p(A) and p(B).

Figure 5.8 shows an example of the feedforward pathway from an s node to the choice probabilities as in figure 5.2. Working backwards from the choice probabilities as before, the probability of choosing label *l* is represented using a similar function to that used in equation 5.1 for the previous model.

$$p(l) = \frac{e^{g_l \sum_c o_{cl}}}{\sum_l e^{g_l \sum_c o_{cl}}} \qquad (5.7).$$

The parameter g_l is a fixed gain on the choice function, sometimes equated with the 'decisiveness' of the network (e.g. Erickson & Kruschke, 1998). The introduction of this parameter to this model is principally to enhance asymptotic performance. As with the IMAW model, association weights between s nodes and c nodes cannot exceed an absolute value of one. *Unlike* the IMAW model, the weights in the c nodes, by which module outputs are multiplied, cannot exceed one. As such, the contribution of any channel to the

decision process cannot exceed an absolute value of one. The gain parameter on the choice function therefore enhances asymptotic performance. This function was fulfilled in the IMAW model, dynamically, by θ weights that could exceed an absolute value of one.

Module contribution is determined by the following function,

$$o_{cl} = \alpha_{cl} \sum_{s} o_{sc} \qquad (5.8).$$

The calculation of the output from a source node s to its channel node c, or o_{sc}, is determined in the same way as for the IMAW model (equation 5.3). The difference between this and the previous model is that the summed, weighted, source node output for a module is multiplied by a value α, which is derived from the adaptive weight θ according to the following logistic function;

$$\alpha_{cl} = \frac{1}{1 + e^{-g_\theta \theta_{cl}}} \qquad (5.9).$$

Because the module associabilities are affected independently of one another by positive and negative signals, according to their relative contribution to the category decision, there is a requirement to prevent associability from falling below zero. 'Negative associability' would appear to have no meaning in this context and using this function obviates the need to 'clip' associabilities at either zero or one. The associability, α, in the current model is the 'expression', in transmission, of the weight θ. The parameter g_θ is a gain or sensitivity parameter which determines how rapidly α changes in relation to changes in θ.

5.2.1.2 Weight update functions

The association weight between a source node s and its channel node c is updated at the end of each trial according to the following function,

$$\Delta w_{sc} = (a_l - o_{sc}) a_s \alpha_{cl} \lambda_w \qquad (5.10).$$

As with the IMAW model, the activation of the label node, a_l, is one if the label is present on that trial and minus one if it is absent. The parameter λ_w is a learning rate parameter held constant for the duration of the simulation.

The update of associability parameters is effected by the alteration of the θ weight that controls them. The update of an associability weight between a channel node C and label node *l*, θ_{Cl} takes place after each trial as follows;

$$\Delta\theta_{Cl} = (a'_l - p(l))\left[\sum_c (k_{Cl} - k_{cl})\alpha_{cl}\right]\lambda_\theta \qquad (5.11).$$

In this case a'_l is the binary representation of label *l*'s activation and, as such, is zero if the label is absent and one if it is present. This function makes use of the overall error of the system with respect to its prediction of the category label. Because this prediction is expressed in terms of a choice probability, the binary representation may be regarded as the probability that the label occurred. The discrepancy is, in this case, that between the probability that the network responds with, for example, A and the probability that the label was, in fact, A. The learning rate constant in this function is, as with equation 5.5, represented by λ_θ.

The second part of equation 5.11 is the sum of the differences between the contribution made by *this* channel, C*l*, towards a correct prediction of the label and the contribution made by other channels, c*l*. These differences are each multiplied by the associability of the 'other' channels prior to summing. This is because it seems appropriate to limit the contributions of channels to the relative effectiveness calculation when those channels are already known to be irrelevant. The measure of contribution to the correct prediction of the label, *l*, by a channel, c, known as k_{cl} is evaluated as follows;

$$k_{cl} = \left|a_l + \sum_s o_{sc}\right| - |a_l| \qquad (5.12).$$

This function will produce negative values if the channel is guessing the wrong way and positive ones if it is guessing the right way. Their magnitude will be the same as the summed source output for the module.

5.2.2. The experimental simulation

The manner of simulation was identical to that carried out for the previous model. Each category structure was run through for sixteen blocks, twenty times with randomised order of input presentation for each simulation.

The parameters used for the simulations were g_l=2, g_θ=2, λ_w=0.1 and λ_θ=0.5. The initial value of all θ weights was set to -0.25 (yielding an initial α value of 0.3775). While these values were not optimised to enhance fit to experimental data, their values could make a difference to the order of difficulty. Higher values of λ_θ tended to enhance the learning rate for the type VI structure relative to that for the types III to V. In this case the

fully valid three-dimensional channel would tend to become dominant in processing much more quickly in the types III to V structures, reducing the positive effects of the semi-valid channels early in training.

The initialisation of the θ weights at a higher value, such as zero, tended to reduce the superiority of the type II structure over the types III to V, particularly early in training. This was a result, as in the previous model, of enhanced capacity for the multiple partially valid channels in the types III to V structures, which makes learning in this model 'look' more like learning in the standard configural-cue model.

5.2.3. Simulation results and discussion

While figures 5.9 and 5.10 show that the model is capable of producing a qualitative fit to the human data, closer analysis reveals that, like the IMAW model, there are problems with the method by which it achieves this.

As can be seen from figure 5.9, the separation of the type VI learning curve from that of the types III to V is somewhat narrow. Examination of figure 5.11 illustrates the principal reason for this in that the type III to V category structures tend to be mostly learnt, by asymptote, in terms of the three-dimensional channel.

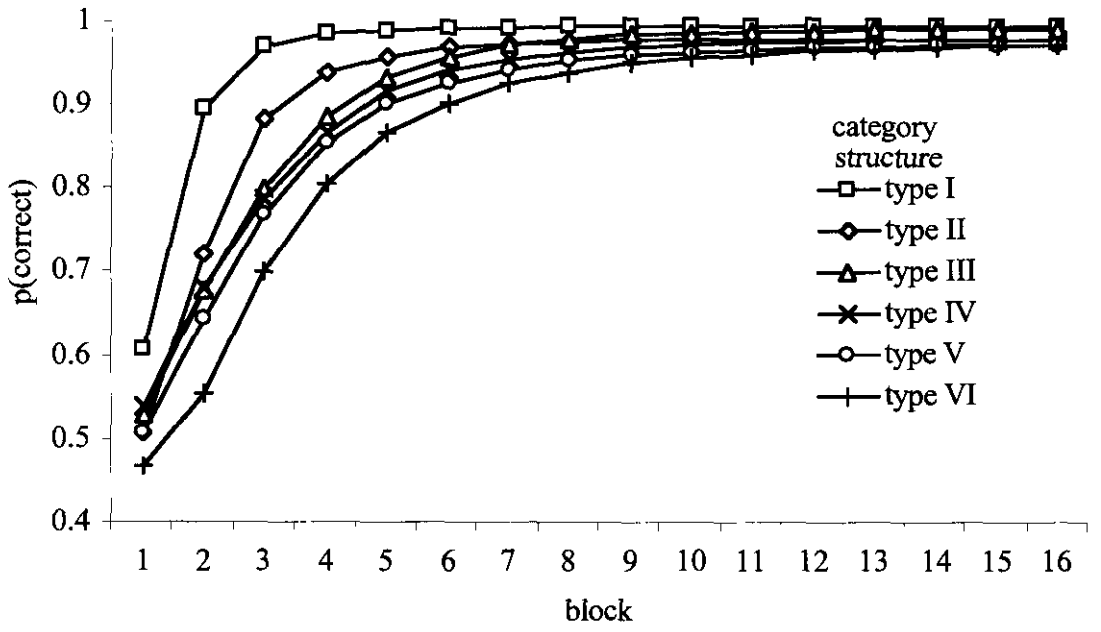

Figure 5.9: Average probability of correct output per block for the relative associability approach for each of the six category structures of Shepard *et al.* (1961). Averages are based on 20 runs through the 16 blocks, on each structure with randomised order of input patterns.

As with the IMAW model, for these tasks the differing logical status of the individual members tends to result in a distribution of response strength across several modules. For different stimuli, different numbers of modules are capable of offering reliable predictions of the category label. Because the θ weight update function changes weights based on the *relative* contribution of modules, weighted by a global teacher signal value, this leads to an asymmetrical pattern of learning in the modules.
Presentation of central members of these categories is likely to lead to higher response probabilities and, consequently, a smaller global teacher parameter affecting the θ weight updates. Because each module is learning according to a local scheme, the level of difference between the output from the fully valid three-dimensional module and that from the semi-valid two-dimensional modules is likely to be fairly low as well.

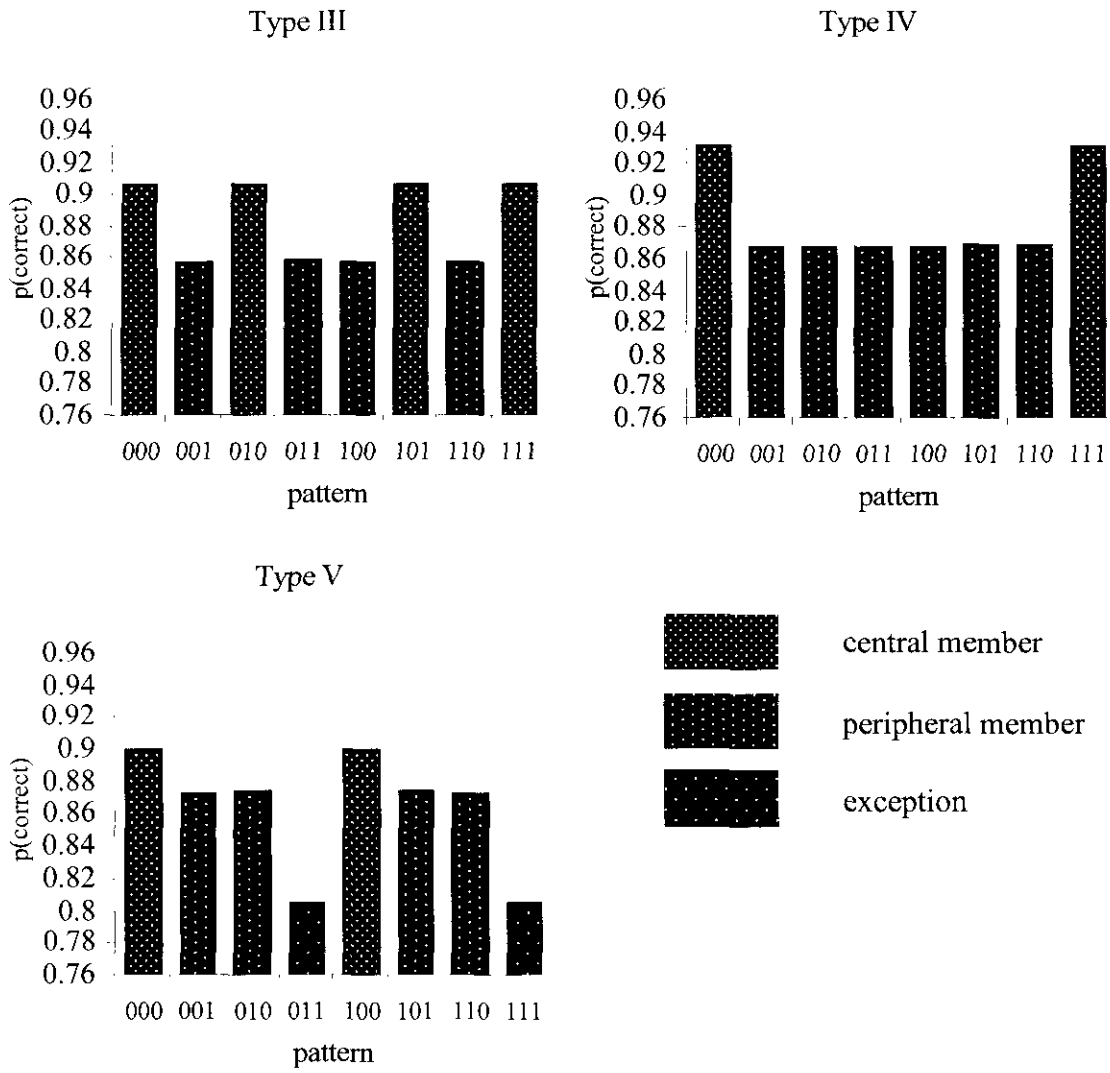

Figure 5.10: Performance of the model on individual patterns for category structures III to V. Performance is indexed in terms of the average probability of correct responding across the entire 16 blocks of the simulation. See figure 2.2 for illustrations of the relationships between these three types of member within each category structure.

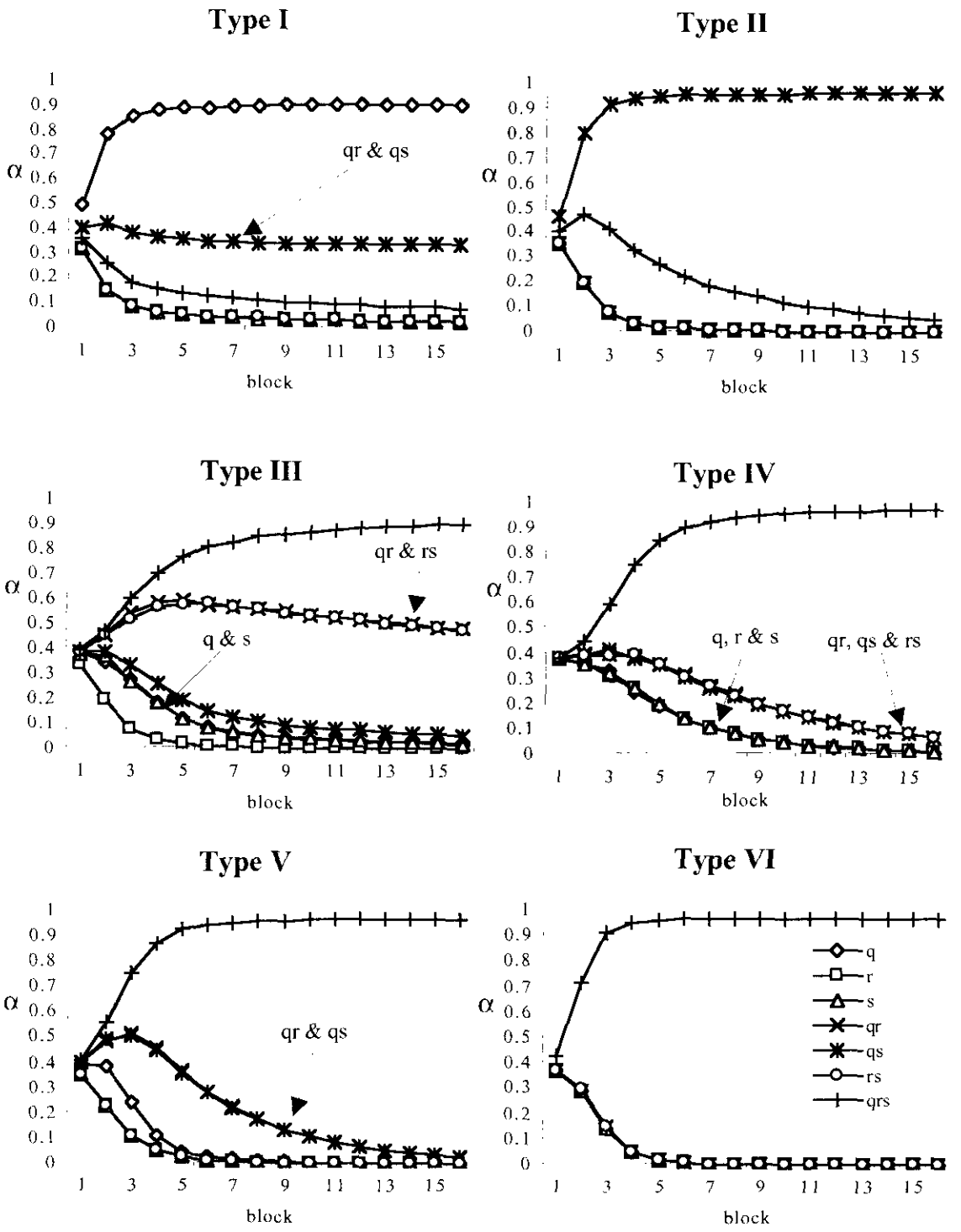

Figure 5.11: Development of α parameters in the model per block on the six category structures, averaged across the twenty simulations per structure. The key shown in the type VI structure applies to all graphs

When peripheral stimuli are presented the global teacher signal is likely to be higher, making the size of θ weight updates larger. In addition, at least one of the two-dimensional semi-valid modules will actually be contributing nothing to the response strength, meaning that the negative signal that it will receive is likely to be quite large (its contribution being much less than that of other modules).

The net result is that update signals during peripheral stimulus presentations are larger than those that occur during central stimulus presentations. For the type V structure, they are largest of all when the exception category members are presented. These signals will generally favour the three-dimensional module which will be providing valid output on all of these trials. Trials where the three-dimensional module is less likely to be providing as high a contribution as lower dimensionality modules are attended by lower global teacher signals. As such, in general, the θ weights for the three-dimensional module increase at a higher rate than they decrease, whereas for lower dimensionality modules the opposite pattern obtains.

Note that for the type III structure, figure 5.11 shows the θ weight for the qs module decreasing faster and earlier than that for the qr and rs module. Examination of this category structure, in figure 2.2, shows that the qs module has two fully valid sources but, in the case of this structure, this module is only ever valid when a central stimulus is presented. Because of this the qs module will not have trials where it is generally the most valid module, as such its θ weights generally decrease.

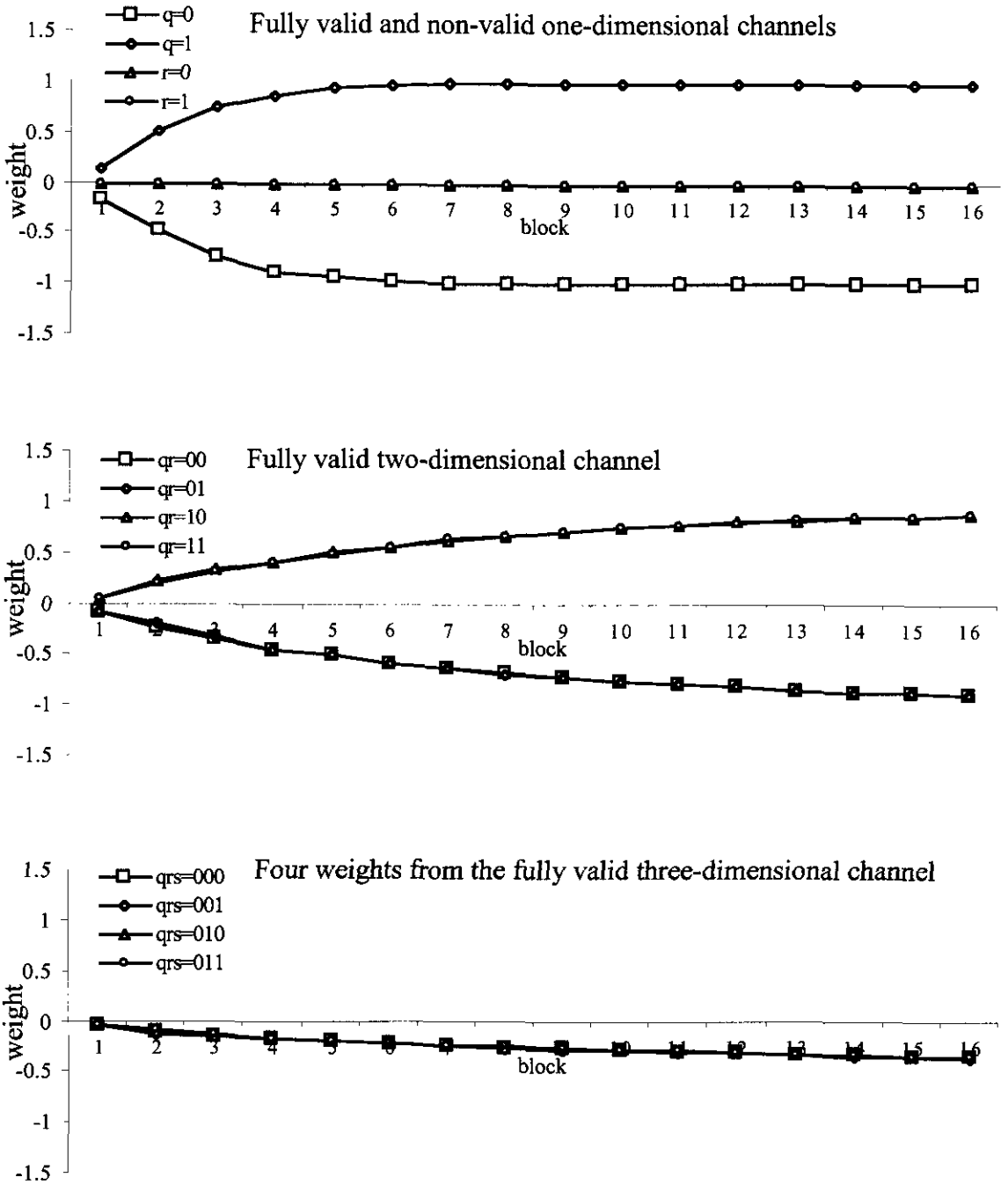

Figure 5.12: Examples of development of source node to channel node (category B) weights across the sixteen blocks of a typical run through the type I category structure.

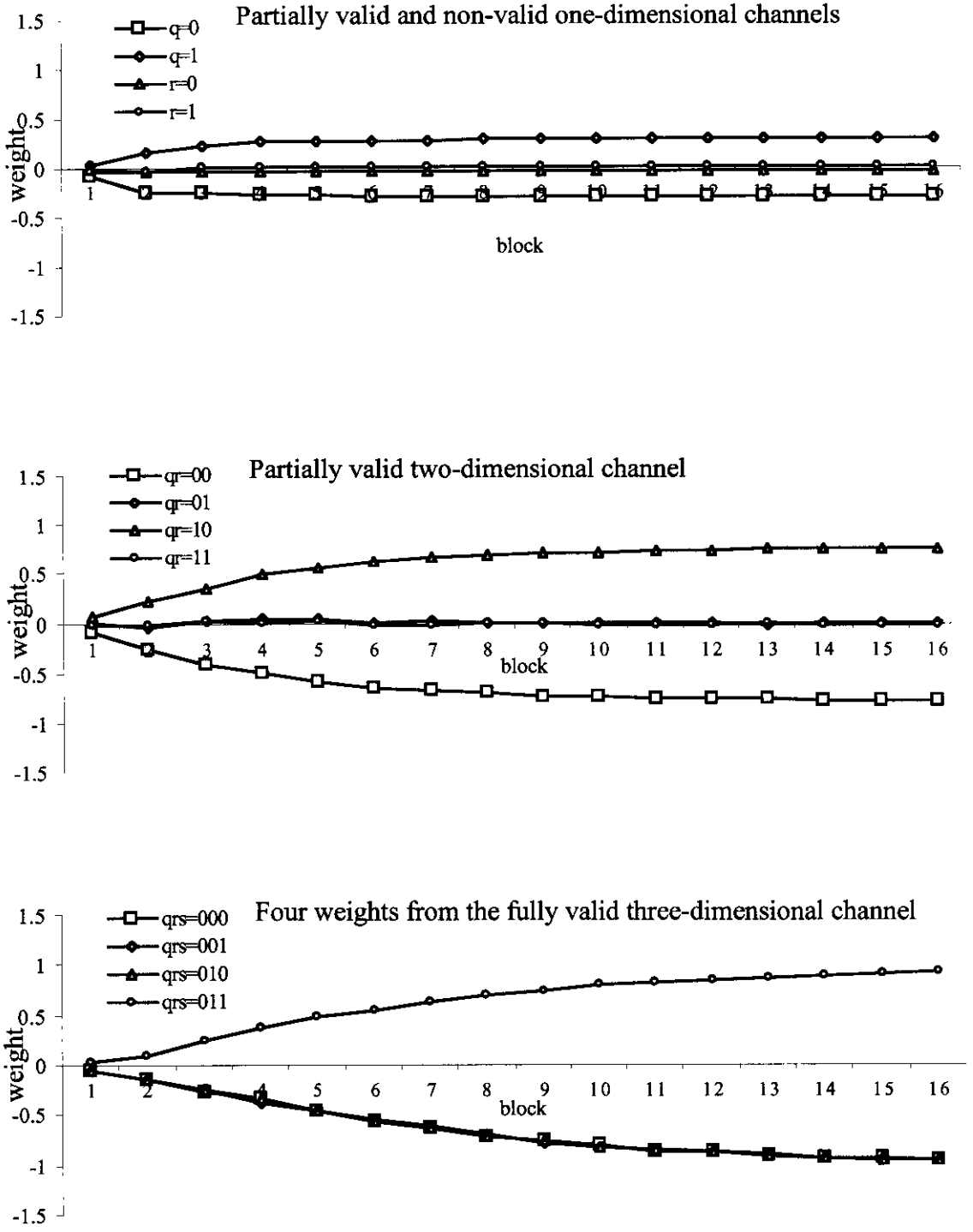

Figure 5.13: Examples of development of source node to channel node (category B) weights across the sixteen blocks of a typical run through the type V category structure.

The early gains in the θ weights for the partially valid two-dimensional modules shown in figure 5.11 are attributable to the higher frequency of the sources in these modules. As the three-dimensional module's θ weights increase, the values of its association weights 'catch' up with those of the sources in these partially valid modules. This is shown in figure 5.13 for the type V structure. It occurs, in this case, at about block 4, which is about the point at which the θ weights for these partially valid channels begin to decrease (figure 5.11).

Unfortunately, removing the global teacher signal parameter from the θ weight learning function tends to narrow the gap between type VI and types III to V curves still further. The reason for this is that once the association weights in the three-dimensional module have caught up with those in the partially valid two-dimensional channels, the θ weights for these channels decrease at a much faster rate. In this case, the fact that error is being reduced by the contribution of the other channels cannot act to slow this decrease. Any contribution being made by these partially valid channels early in learning is subsequently removed by the decrease in the associability weights, such that the only significant contribution, later in learning, comes from the fully valid three-dimensional module.

Another problem with the model is associated with the development of associative strength for sources in redundant, relevant modules. This is shown clearly in figure 5.12 which illustrates the development of associative weights in the type I category structure. Comparing this figure with figure 5.11 reveals that because the overall error regarding this task is reduced so quickly, the θ weight for the two-dimensional modules do not have time to reduce to a low enough level to seriously attenuate associative learning in their modules. The result is that these associative weights continue to increase.

This learning occurs, albeit to a lesser extent, for associative weights in the three-dimensional module. Although not shown here, it is also observed for the three-dimensional module in the type II task. As figure 5.11 shows, it takes a while for the θ weight for the three-dimensional module to decrease in this task. During this time learning of associations within this module will continue.

This approach would appear to inherit some of the problems of the IMAW model given previously. Like this model it may have difficulty with compound-component

discriminations without some explicit representation of the absence of a component. In this case the problem is slightly different. Due to the local learning rule in the modules and the fact that associability weights cannot exceed one, the compound's associative strength cannot exceed that of the components.

Like the previous model, this model would have some difficulties representing blocking of conditioning. Similar to the IMAW model, it is not easy to specify what the starting value for an associability weight should be. When the compound is presented in the blocking paradigm, for example, if the new associability weight for the two-dimensional module is greater than zero, associative learning will occur in the new module. Furthermore, because the global error should be close to zero on account of the previous training with the component, there will be little or no change to the value of the associability weight, despite its lower validity on each trial.

As suggested for the IMAW model, one could specify some rule which set the initial value of associability weights as a function of the discrepancy that obtained on the trial on which they appeared. As with the suggestion in the context of the IMAW model, however, this would require one to specify a set of rules that relate associability weights to error throughout the course of training. While using error information in this way may facilitate increased generalisability of the approach, it remains an area for future research.

5.3. General discussion

The two models presented in this chapter demonstrate that modularly organised configural-cue architectures are capable of modelling the data reported by Shepard *et al.* (1961) and Nosofsky *et al.* (1994). As suggested in the previous chapter, the difficulties experienced by Nosofsky *et al.* (*ibid.*), when testing their modular DALR variant of the configural-cue model, can be overcome by the use of alternative measures of average module performance.

There are two, related, key aspects to both models presented above. The first is that the sources of the configural-cue representation are organised into 'spatial' modules. The second is that the contribution of each module to the decision process is gated by an adaptive weight which, in some way, reflects the *average* validity of the spatial module with respect to the category structure being learnt. This weight also controls the maximum rate at which representations within the module acquire associative strength.

This spatial modular structure may be related to rule-based accounts of the categorisation process such as that advanced by Shepard *et al.* (1961) and Feldman (2000) to account for the relative difficulty of category structures.

Each module may be conceptualised as a 'candidate' rule. Its dimensionality is related to the number of clauses required to define the rule that it may represent. The dimensionality of the module, with respect to these tasks at least, determines the frequency with which each representation within that module will be instantiated. This in turn determines the rate at which the module will 'learn'. This much alone provides the reason why type I is easier to learn than type II which is, in turn easier than type VI. Low dimensionality rules are learnt more quickly than high dimensionality rules because their conditions are tested and evaluated more frequently.

The two models presented in this chapter represent two quite different ways of accounting for the relative difficulties of the Shepard *et al.* (1961) category learning tasks. The two distinct algorithms for updating modular weights reflect the two different types of information used by the models. The independent modular associability weights model (IMAW) bases its account on, firstly, the redundancy of the modules' contributions with respect to the task. Secondly, the model relies on an assumption that a module's output to the decision process is gated by some value reflecting the average validity of the module's contribution to the decision process. This assumption is generally architectural, in that it can be modelled in terms of an independent modular weight which tracks the average performance of the module on each occurrence of a label.

The relative modular associability weights model (RMAW) is somewhat different in that it offers an account based on the relative merits of each module's reinforcement schedule in relation to each task. Both models are fundamentally dependent on the idea that sources or representations are organised in a way which is based on the 'space' within which they occur.

The redundancy, upon which the IMAW model is reliant, is inherent in the modular configural-cue representation. Each dimensional value, or configuration of values, is represented 2^{o-d} times, where o is the dimensionality of the whole object and d is the dimensionality of the value or configuration in question. Simpler rules get learnt faster, according to IMAW, not only because their lowest dimensionality representation gets

tested and evaluated more frequently than higher dimensionality rules. They are also faster because they can be represented or duplicated by more of the modules in the architecture.

The average validity of each module, for IMAW, is measured in terms of a weight which reflects how well the module's output predicts the category label. Because this weight tracks the output from the module across all trials, its maximum value will generally be less than one if the module is not fully valid for the task. Because the hyperbolic tangent of the weight (multiplied by plus or minus one, depending on the label) is compared with the output of the module, the maximum value of the modular weight is greater than one if the module is fully diagnostic.

The effect of this scheme is such that partially valid modules are substantially penalised, relative to fully valid modules, both in terms of the contributions they may make to the decision process, and in terms of the learning rates of their sources. The key to 'making' this modular scheme reflect the observed order of task difficulty lies in controlling the contributions of partially valid modules. The combined contributions of up to three, partially valid two-dimensional channels, and up to three partially valid one-dimensional modules, must not exceed the ongoing contributions of a single, fully valid two-dimensional module. This scheme would appear to meet that criterion such that module redundancy may be used as an index of task difficulty.

The RMAW model makes use of a different 'type' of information to represent the observed order of task difficulty. In this case associability weights are controlled by the differences between the reinforcement schedules of each module. Modules begin with equal, intermediate-value, associability weights. An associability weight will increase if the module it pertains to has a clearly superior reinforcement schedule to the other modules.

This superiority may be manifested between fully valid modules in terms of the relative frequencies of their sources. A one-dimensional fully valid module has a superior reinforcement schedule to a fully valid two-dimensional module. It is this relationship which results in the correct I, II, VI ordering of difficulty.

The reason why it penalises the types III to V tasks relative to the type II task is that, early in learning, there is no channel which is clearly best able to represent the category learning task across all trials. This tends to result in a persistence of intermediate-

value associability weights for each of the partially relevant modules and, as a result, associative learning in these modules is attenuated, regardless of the individual validity of sources.

This usage of the modular architecture highlights an important aspect of its relationship with dimensional attention models. When the modularity of a connectionist network is spatial, as it is with the two models presented above, each module may, effectively, represent a particular dimensional attention pattern.

This is most evident with the RMAW model, where there is an element of competition between each module in terms of the extent each contributes to the decision process, and the rate at which its sources learn. In this case the model is effectively selecting which of its modules is learning the category structure the fastest. For the types I, II, and VI structures the modules selected (or the associability weights of each module) correspond to the adoption of a dimensional attention strategy similar to that shown by the GCM (Nosofsky, 1984) for these three tasks (see table 3.6).

While attention is not specifically dimensional in this model, the effect is the same. Information about dimensions which are not essential to the task is not used in the performance of the task. Rather than select the dimensions themselves, as with ALCOVE, the relative modular associability model selects from the range of 'spaces' which would result from particular patterns of dimensional attention.

As discussed in sections 5.1.4 and 5.2.3, however, the modifications made to the configural-cue architecture, for the two models, may have resulted in these models losing important aspects of the functionality of the basic configural-cue model. Without further modifications, for example, neither model seems like it would be capable of representing blocking. This particular loss of functionality is primarily due to the use of associative learning rules that do not take into account the combined associative strength of all active representations. The attention learning methods used by the model do not appear to offer alternative explanations for the data predicted by the basic configural-cue model.

In the course of allowing the configural-cue model to make comparable predictions to models such as ALCOVE (Kruschke, 1992), it would appear that the resultant models have acquired some of ALCOVE's shortcomings. In particular this relates to the models' ability to describe certain learning effects, such as compound-component discriminations.

Another criticism of the models is that their abilty to represent the Shepard *et al.* (1961) data appears to be highly dependent on the initial values given to their modular associability weights. If these initial values are too high, then the task difficulties tend to approach those observed in the basic configural-cue model (Gluck & Bower, 1988b, Nosofsky *et al.*, 1994). If the values are initialised too low then, owing to the dependence of associative learning on the magnitude of these weights, early learning is severely attenuated. The intermediate initial values of these weights seem to be required to allow the effects of average module validity to be manifested at as early a stage in learning as possible. There is, however, no principled reason for their initial values, other than the fact that they produce the desired results for the task.

Whether these values would allow the models to simulate human performance on other tasks may require further testing. It seems likely, however, that more sophisticated rules relating initial modular weight values to error or uncertainty, as suggested above, may be required to allow these models to simulate basic associative learning capacities.

In summary, while these models produce a qualitative fit to the data, this does not appear to be all that robust. There does not appear to be any principled reason why the parameters required to achieve even this qualitative fit should be set as they are in the tests above. In addition the models appear to have lost much of the functionality of the basic configural-cue model. The next chapter details models which incorporate dimensional attention into the configural-cue form of representation. As will be discussed, this approach seems to be more robust with regards to its ability to produce the required qualitative fit. In addition, because these models are not modular they can incorporate global error parameters in their learning rules. This enables preservation of much of the functionality of the basic configural-cue model.

Chapter 6: Modelling Shepard, Hovland, and Jenkins (1961) using configural-cue networks with dimensional attention

The final model presented in chapter 4 employed a method for representing differential *dimensional* contributions to the activation of representations. This approach made use of a sequentially represented process of dimensional sampling with representations or detectors activated only by consecutive sampling of their component dimensions.

In this approach, the sampling process is representable in terms of a Markov process. The simple model in chapter 4 used a system where the probability of a particular dimension being sampled at time t+1 is only a function of the sampling probabilities for each dimension and the dimension sampled at time t.

In this approach, the sampling process is dependent on the vector of initial, modifiable sampling probabilities, **p**=(p(q), p(r), p(s)) and the matrix of transition probabilities **P**={p(j| i)}. The probability p(j| i) represents the following conditional probability $p(X_{t+1}=j| X_t=i)$, which is the probability that the dimension sampled, X, at t+1 will be j given that the dimension sampled at t is i. In addition, the probability of a dimension being sampled on consecutive time steps was affected by a global parameter relating to the overall uncertainty of the sampling decision process.

In this chapter, three models will be presented. One of these is a direct attempt to model the transmission rate model with dimensional attention, presented in chapter 4. In this model, the parameters altering at the end of each trial are the initial sampling probabilities in **p**. A second model represents the transition probabilities more directly and attempts to use a back-propagation of error algorithm to alter the matrix **P** between trials. A third model, which is based on the modifiable transition matrix approach, attempts to address some of the conceptual shortcomings of the other two models.

6.1. Alteration of sampling probabilities by back-propagation of error: The Adaptive Sampling Probabilities (ASP) model

The basic architecture of the model is shown in figure 6.1, illustrating the relationships between the various components. As with the transmission rate model in section 4.3.3, and unlike the two models presented in chapter 5, the source nodes are not explicitly grouped in a modular way. They are shown as groups in figure 6.1 to illustrate their common dimensional input.

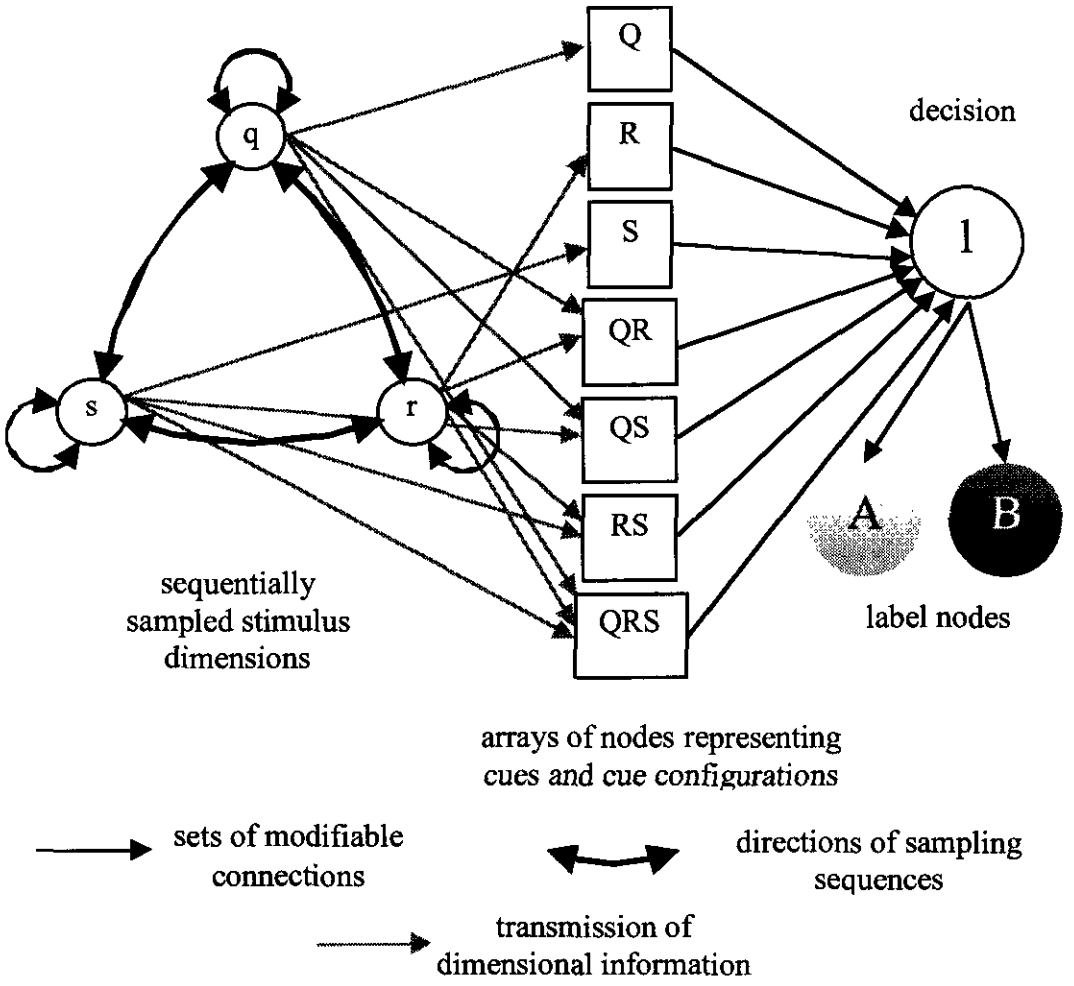

Figure 6.1: Adaptive sampling probabilities (ASP) model showing dimensions, channels (arrays of nodes), label nodes, and the relations between them. Note that each array contains a number of 'source' nodes: 2, 4, and 8, for 1, 2, and 3-dimensional arrays respectively. Each of these nodes is connected with the decision node.

6.1.1 Feedforward functions

Unlike the models presented in chapter 5, this model only uses one output weight per source node. This is illustrated in figure 6.1 by a separate 'decision' node labelled l. Output from the source nodes is summed and passed through a logistic to determine the choice probabilities. The probability that the model selects A, or p(A) is given by the following equation,

$$p(A) = \frac{1}{1 + e^{-g\sum_{c} o_{cl}}} \qquad (6.1).$$

Here g is a gain or confidence parameter, which modulates the slope of the logistic, p(B) is given by 1-p(A). The output from a given node array or channel, c, to the decision node l, or o_{cl}, is a function of the probability that the nodes in that channel are active and the maximum contribution those nodes could make, if they were active.

$$o_{cl} = p(c)k_{cl} \qquad (6.2),$$

where k_{cl} is the maximum contribution to the decision node given by,

$$k_{cl} = \sum_{s \in c} a_s w_{sl} \qquad (6.3).$$

As for previous models, a_s has a value of 1 if the cue or cue configuration which the source node s detects is present and zero if it is absent. The output from the source node is, as given in equation 6.2, dependent on the probability that the channel is active on that trial. In the case of a one-dimensional source, this is dependent on the probability that the dimension is sampled on a particular trial. For sources with more than one dimension, this depends on the probability that the component dimensions of the source are sampled consecutively.

6.1.2. Sampling and activation probabilities.

In this model, a channel 'c' is simply a set of source nodes that receive their input from the same dimension or set of dimensions. The probability, p(c), represents the probability that a channel is activated, and thus sending a signal, on a given trial. This is dependent on the dimensions involved in the channel and the probability of them being sampled during a trial. The model of node activation is the same as that used in the transmission based model given in section 4.3.3, with nodes activated according to the scheme illustrated in figure 4.13.

A slight exception to the scheme used in the transmission rate model involves the use of the sampling decision entropy parameter, H(D), in calculating the transition probabilities. For the ASP model a parameter, h, was incorporated, which controlled the effect of the decision entropy on the sampling process. Assuming three dimensions i, j, and k, the entry in the matrix corresponding to $p(j_{t+1}| i_t)$ is calculated as follows;

$$p(j_{t+1} \mid i_t) = \frac{p(j)}{p(j) + \left(p(i)\left(1 - \left(H(D)h\right)\right)\right) + p(k)} \qquad (6.4).$$

This replaces equation 4.25 for the determination of transition probabilities for the model. The parameter h controls the effect of decision entropy (determined using equation 4.23) on recurrent sampling probabilities. The lower its value, the lower the effect that the entropy has on the process.

This 'across-trial' average chance of activation does not directly address the method by which sampling is 'translated' into node activation. For the purposes of this model it is assumed that the probability of the sampling event required, across the trial, may be mapped directly onto an activation level at the point of decision and feedback.

The probabilities of node activation are determined using identical equations to those used in chapter 4, section 4.3.3.2, for determining the transition matrix and the channel activation probabilities. In this case a node is active according to the probability that its channel is active, or p(c).

6.1.3. Updating weights and sampling probabilities

Updating the associative weights w_{sl} is achieved at the end of each trial using a variant of Rescorla and Wagner's (1972) learning rule.

$$\Delta w_{sl} = (\delta - p(l))a_s p(c)\lambda_w \qquad (6.5).$$

The change in the weight is, here, based on the difference between the teacher signal δ, which equals one if the label was A and zero if the label was B, and the output of the network passed through the logistic function as in equation 6.1. This measure is multiplied by the activation of the node and the probability that the node's channel, c was activated on that particular trial. The learning rate parameter for these weights is given by λ_w. The weight change is added to the weight for the next trial.

The model uses the difference between the teacher signal and the *probability* of selecting a particular label, rather than the raw output, to determine the weight update. This choice was motivated by theoretical and practical considerations.

The theoretical considerations are those discussed in the context of Kruschke's (1992) use of 'humble teachers' in section 3.3.3.2.1. With the nominal feedback used in category learning experiments, the goal of associative learning may be best described as one of minimising the uncertainty of the decision process regarding category membership. The magnitude of this uncertainty is not really indexed by the raw measure of associative strength delivered to a single alternative in the decision process.

The best measure available will be derived from the interaction of strength for each alternative, in the form of the response probabilities themselves. The difference between the probability of a response and a measure of whether that response was correct or not (given by the label feedback) would thus seem an appropriate candidate to describe the magnitude and direction of learning which might be required.

On a more practical level, asymptotic performance of the model may be compromised by the use of the raw output measure when the error is going to be used to determine a back-propagated signal. This is particularly the case when the weights affected by the back-propagated signal are normalised subsequent to their update. Even slight negative signals owing to an 'excess' of output will be passed back maximally to the most relevant dimension due to the larger associative weights which are dependent on relevant dimensions.

The implications of the use of this variant will be discussed more in the next chapter, when its generalisation to a situation with more than two alternatives is required.

Updating the dimensional sampling probabilities is based on the back-propagation of error via the channels dependent on each dimension. The increment for a dimension d is as follows,

$$\Delta p(d) = \sum_{d \in c} (\delta - p(l)) o_{cl} \lambda_p \qquad (6.6),$$

where λ_p is an update rate constant. Increments are added to the probabilities and then normalised to sum to unity. If p(d) plus its increment is less than 0.0001 then p(d)+Δp(d) is clipped at 0.0001 prior to normalisation.

6.1.4. Results from the experimental simulation and discussion

As with the previous models, each category structure was run through for sixteen blocks of sixteen trials, twenty times each with a different randomised order of presentation.

6.1.4.1. Overall performance results

Figure 6.2 shows the mean probability of selecting the correct category per block of training across the twenty runs. The parameters used for this particular set of results were h=0.65, λ_w=1, g=2.5, and λ_p=1.

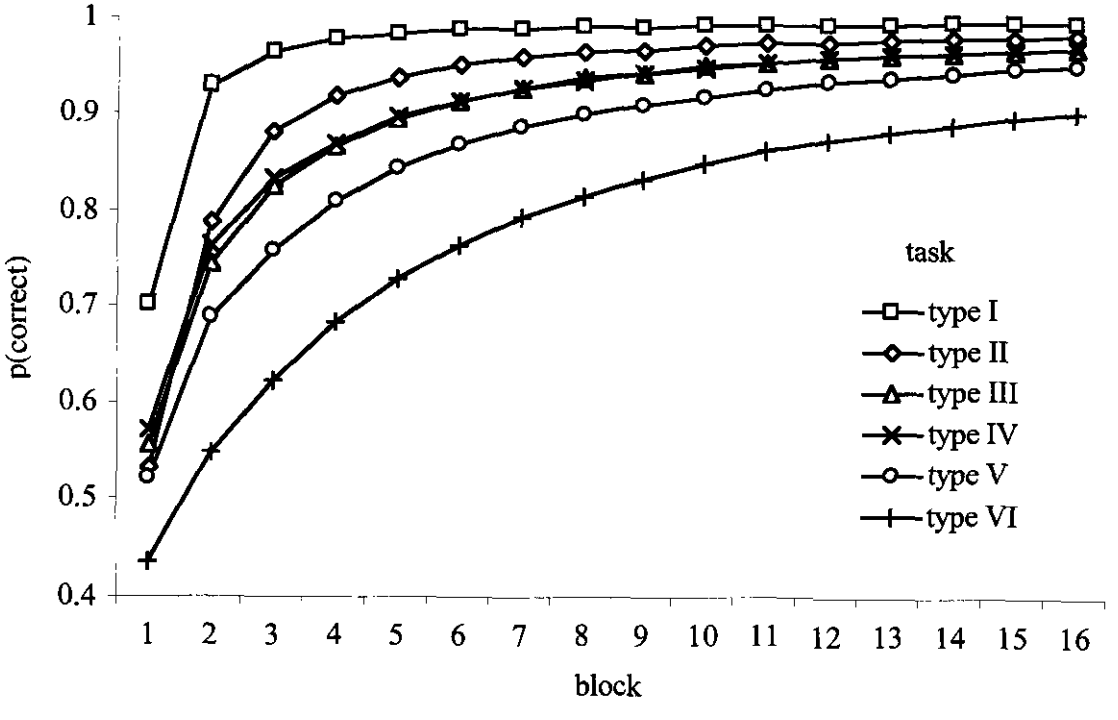

Figure 6.2: Mean p(correct) per block averaged across the twenty runs through each task for the adaptive sampling probabilities (ASP) model.

As with the previous models, this model displays the correct ordering of difficulty for the tasks. The main concerns here are as follows:

1. As with the previous models, there are still problems with the late emergence of clear superiority for the type II structure over the types III to V. The nature of these problems is the same as those suffered by these other models and concerns the early dominance of one-dimensional contribution to output. This is not likely to be alleviated by this particular approach as even were it to sample the 'correct' valid dimensions early on, this will still allow the contribution of nodes representing the non-valid one-dimensional inputs.
2. As with the IMAW approach (but not the RMAW model), there seems to be a trend on the type III to V structures for difficulty to be, approximately, a function of the number of valid sources. The order of these three structures is IV< III < V. Although the differences between the performances are small, they do appear to be consistent. In addition the average difference between the type IV and type V structure would appear to be approximately as great as that between the types III and IV tasks and the type II task.
3. The performance on the type VI task, for the first two blocks, is fairly poor. Average performance on the second block of training is only just above chance level. As with the type II problems, this is likely to be the result of interference from non-valid sources during the early trials. Because the activation of three-dimensional representations is less than that of these 'interfering' representations, this interference is likely to persist for longer than it does on the type II structure.
4. Fairly poor asymptotic performance on all tasks other than the type I structure.

On the positive side, the difference between performance on the type VI and type V task is greater than that displayed by the modular models. The activation functions for the three-dimensional representations required for the type VI structure, are, in this model, dependent on the probability of the component dimensions being sequentially sampled. As with the transmission rate version of this model reported in chapter 4, this makes the activation for nodes a function of their dimensionality and, consequently, 'penalises' the three-dimensional representations most heavily.

Figure 6.3 also shows that the model, correctly, has more trouble with learning peripheral category members than central members. For the type V task, the model finds exception members the most difficult. The weight distributions which underlie this pattern will be discussed below.

6.1.4.2. Sampling probabilities and channel activations

Figure 6.4 illustrates the evolution of the sampling probability vector, **p**, as the model learnt the tasks. As can be seen, the effects of learning are particularly evident in the types I and II tasks, where the model quickly develops high sampling probabilities for the valid q dimension in the type I task, and the configurally valid q and s dimensions in the type II task.

For these two tasks the influence of redundant relevant channels is eliminated fairly quickly. As will be described below, because the model makes use of a global teacher signal for the update of its associative weights, learning in redundant relevant channels is fairly well attenuated. In this case, the use of a variant of the Rescorla-Wagner learning rule does not appear to compromise performance at all. This suggest that the model may be capable of representing blocking of conditioning, whilst at the same time be capable of exhibiting the dimensional attention which appears to be required for the Shepard *et al* (1961) tasks.

There is almost no deviation from the initial pattern of equal sampling probabilities for the type III task, and nothing at all for the type IV task. In the case of the type III task, there is a very slight tendency for the sampling probability of its *non-valid* dimension r to be greater than that for the partially valid dimensions q and s. This will be discussed in more detail below and also in the context of the next model, where this tendency is much more marked. This tendency, it will be noted, is similar to the RMAW model discussed in section 5.3.3. As can be seen from figure 6.4, the sampling probability for the partially valid q dimension in the type V task is slightly higher than the probability of the remaining dimensions.

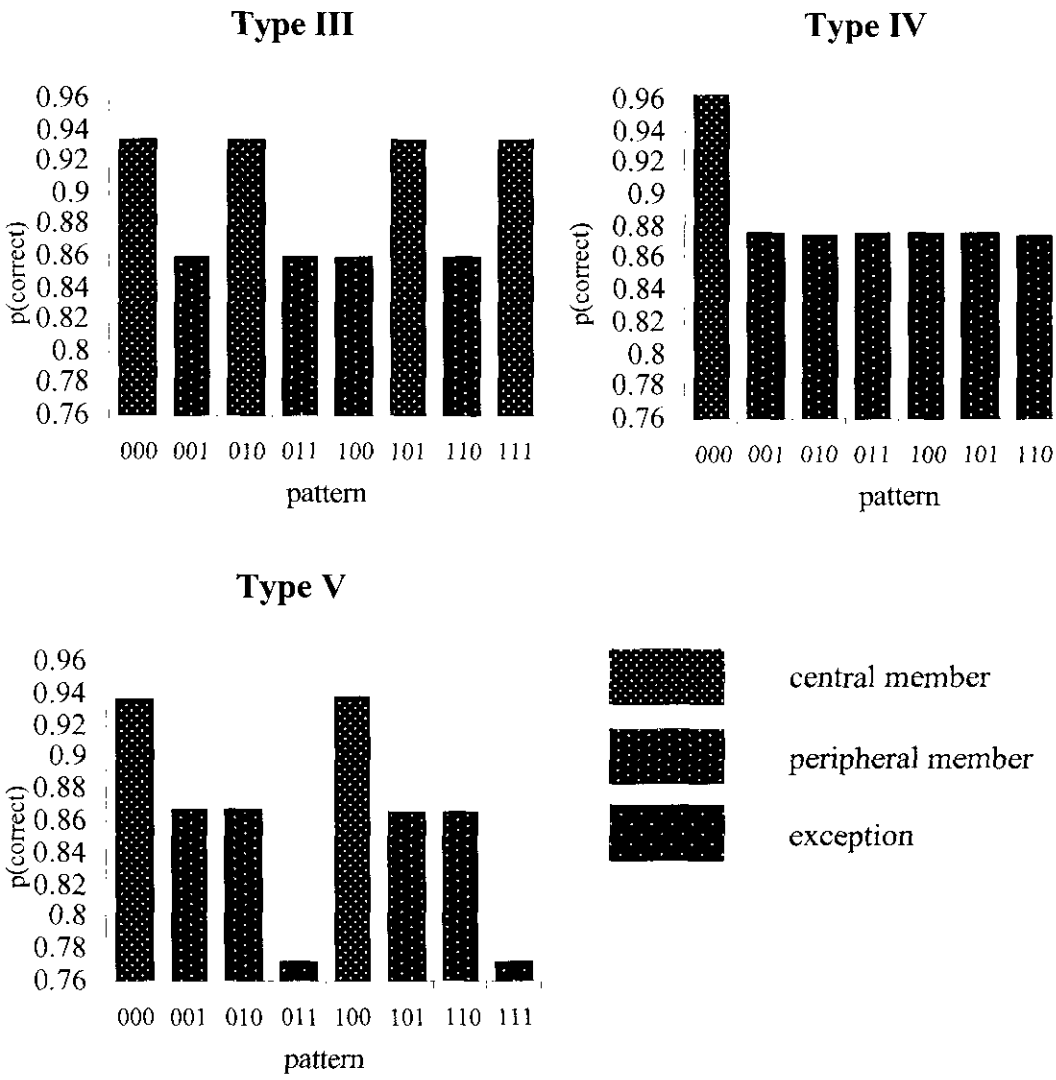

Figure 6.3: Performance of the model on individual patterns for category structures III, IV, and V. Performance is indexed in terms of the average probability of correct responding across the entire 16 blocks of the simulation. Figure 2.2 shows the relationships between these three types of category member in each category structure.

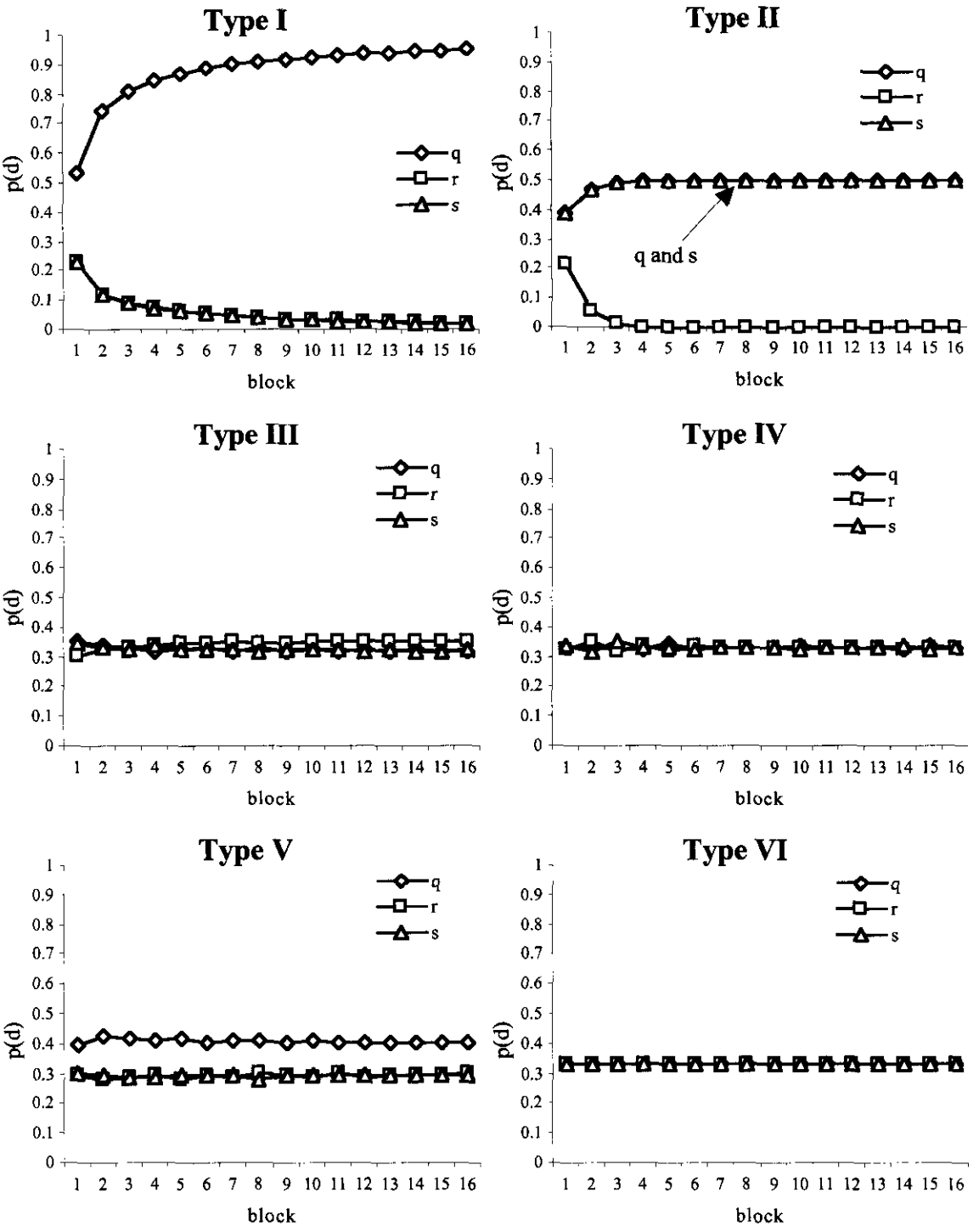

Figure 6.4: Average dimensional sampling probabilities, (vector **p**), per block.

The manifestation of these probabilities in terms of the actual channel activations is shown in figure 6.5. It is important to note that the maximum activation for configural channels, such as the valid QS channel in the type II task, or the QRS channel, is controlled by the parameter h (used in equation 6.4). Increasing the level of this parameter increases the effect of sampling uncertainty on the probability of consecutive sampling of the same dimension.

The QRS channel is particularly affected by this value. If the value of h were one, the probability of QRS activation, given p(q) = p(r) = p(s) would be 0.5. As can be seen from figure 6.5, for h=0.65 this value is approximately 0.36. Altering this parameter tends to increase performance on tasks reliant on configural channels.

Figure 6.5 illustrates that there is very little difference in the activation probabilities for the channels in the types III to V tasks. This, in itself, suggests a reason for why there is such a difference between performance on the types III and IV, and the type V task. As discussed in the context of the IMAW model, there are simply more valid sources for the types III and IV task than for the type V task. The lack of difference between the channel activation probabilities for these tasks means that the overall rate of learning will be, principally, a function of the number of valid sources for each stimulus.

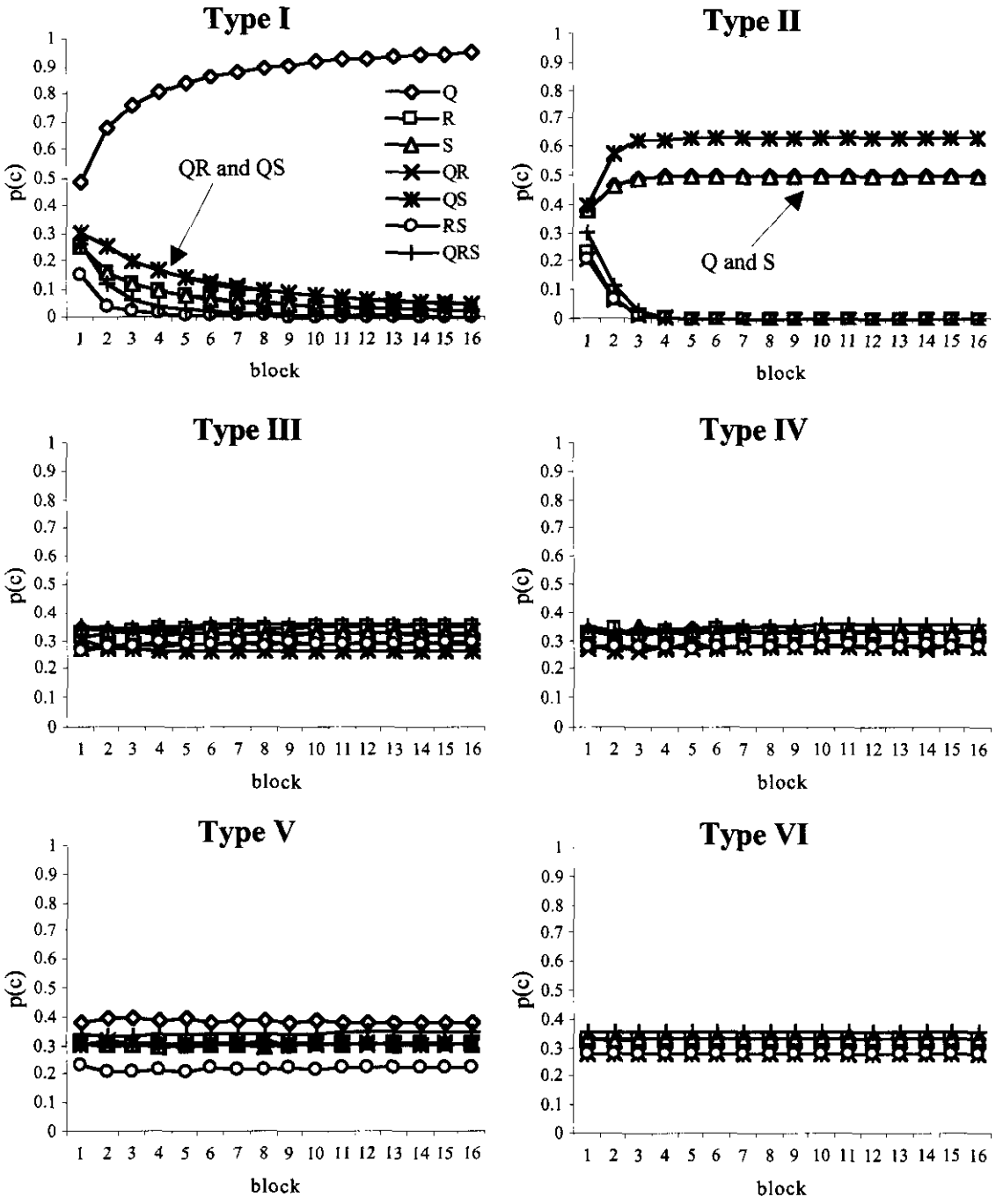

Figure 6.5: Average probability of each channel being active, p(c) on a trial per block for each of the six tasks. The key given for the type I task applies to all graphs.

6.1.4.3. Associative weights

Figures 6.6 to 6.9 show the development of associative weights for some of the channels on some of the category structures. These weights are the weights at the end of each block for one run through the relevant category structure. Note that the order of difficulty on these five runs was the same as that shown in figure 6.2.

Figure 6.6 shows the development of weights in the type I structure. Weights are shown for the valid Q channel, the non-valid R channel, the valid but redundant QR channel, and four weights from the QRS channel. As mentioned above, the model shows considerable attenuation of learning in the redundant relevant higher dimensionality channels.

Weights for the type II structure are not shown here but learning in the redundant relevant QRS channel for this structure was similarly attenuated. A typical magnitude of a QRS weight in this task was about 0.3 compared to asymptotic magnitudes of about 2.5 for the QS channel.

Figure 6.7 shows the associative weights developed for five of the channels on the type III task. These graphs show an interesting property of the interaction between the global teacher signal and the configural representations used. The partially valid QR channel has two valid sources, qr=00 and qr=10, and fairly large weights develop on these connections. The two non-valid sources, in this case, also have weights that are almost as great as those developed on the one-dimensional, partially valid Q sources shown in the top panel of figure 6.7.

This pattern is also shown in the type V task for the QR channel, shown in the centre panel of figure 6.9. A closer analysis of the weights and the tasks reveals that for each of these partially valid channels, when one of the dimensions is independently partially valid and the other is non-valid, weights will develop on the non-valid two-dimensional sources. These weights will be in the opposite direction to the 'rule' promoted by their single partially valid dimensions.

As can be seen from figure 6.7, the qr=01 weight is positive whereas the q=0 weight is negative, the qr=11 weight is negative whereas the q=1 weight is positive. A similar pattern is shown for the weights on the non-valid qr=01 and qr=11 sources in the type V task shown in figure 6.9.

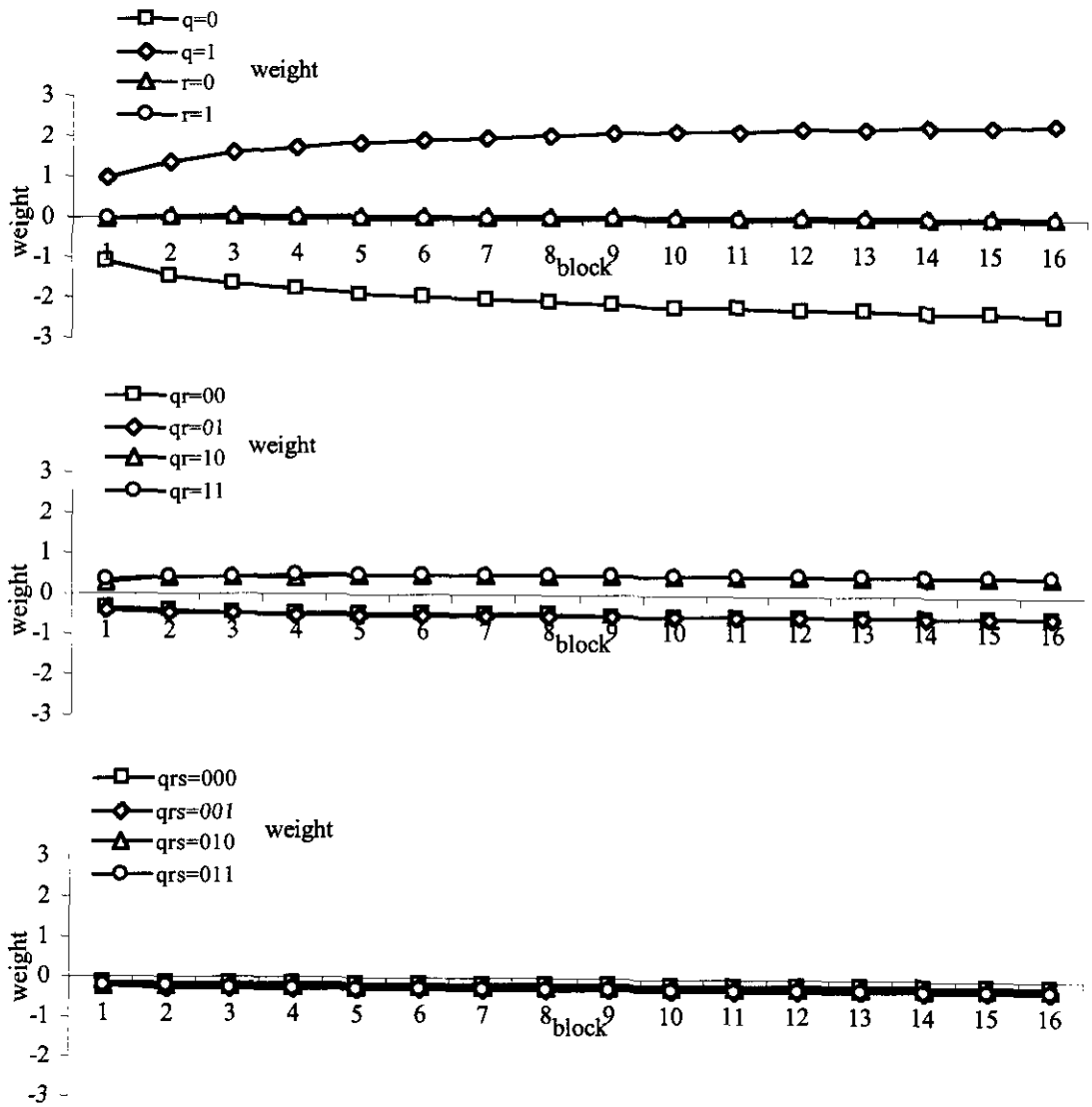

Figure 6.6: Associative weights for the Q, R, and QR channels and four weights from the QRS channel (to category B) at the end of each block of learning for one run through the type I category structure.

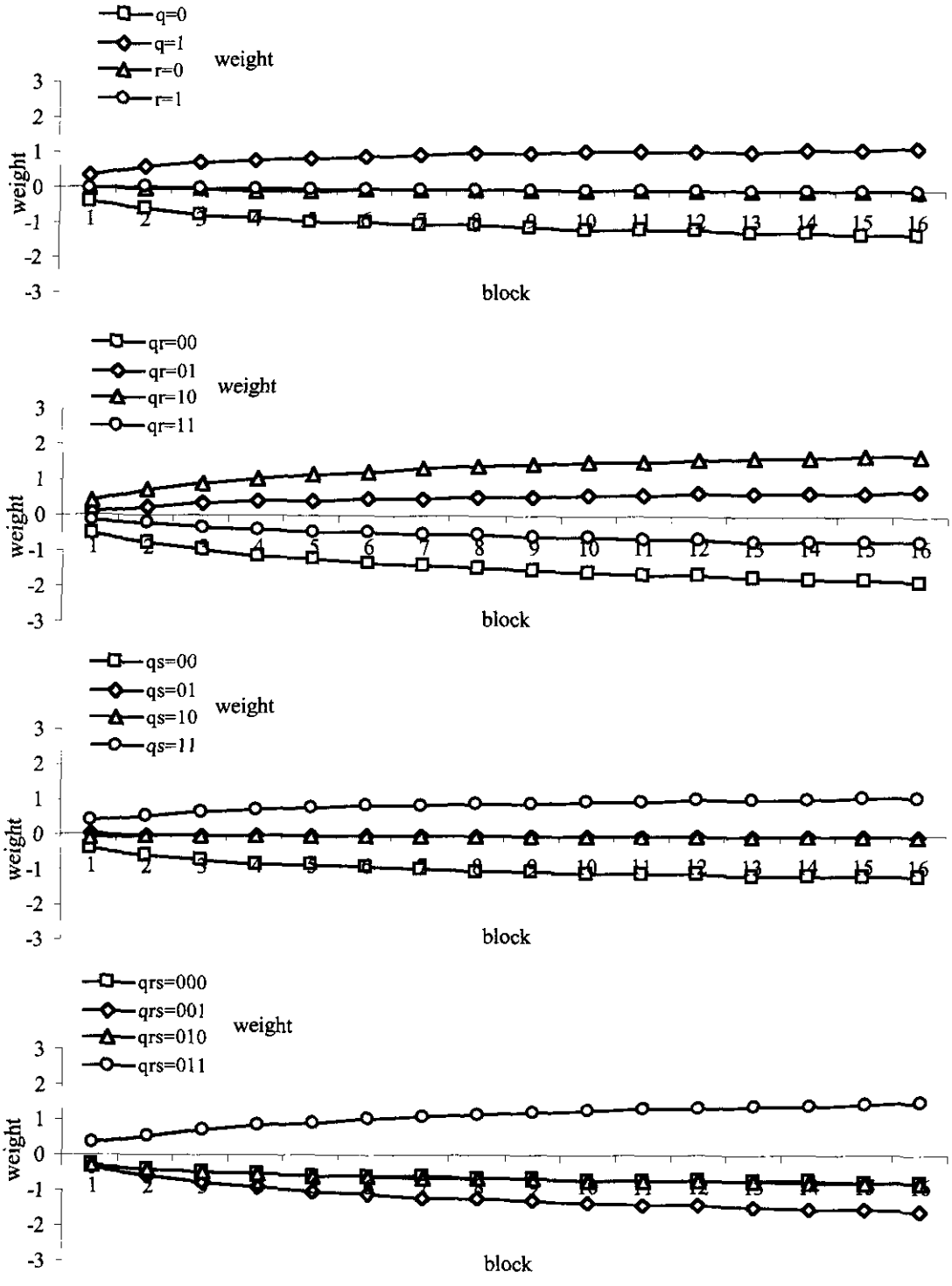

Figure 6.7: Associative weights for the Q, R, QR and QS channels and four weights from the QRS channel (to category B) at the end of each block of learning for one run through the type III category structure.

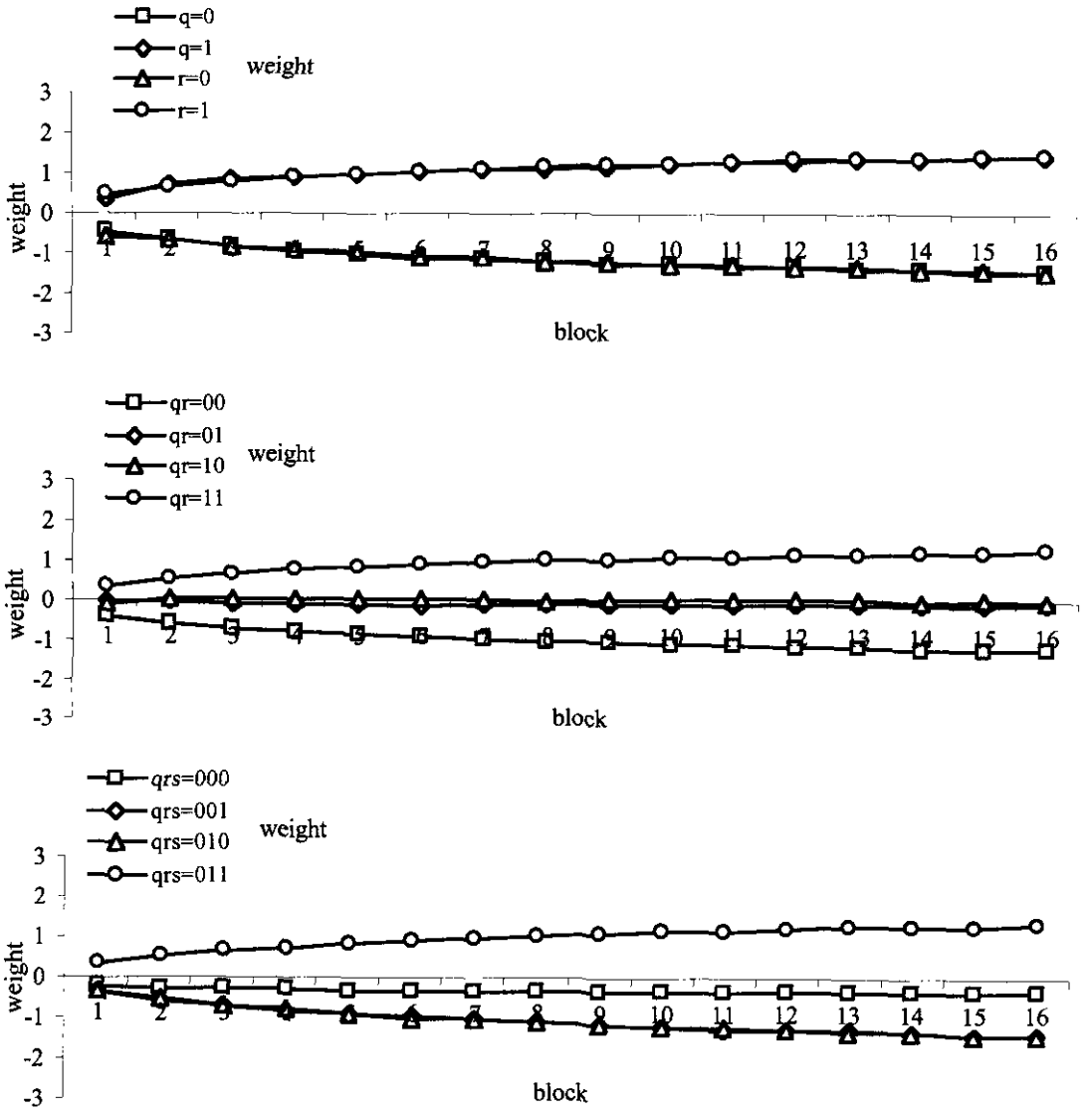

Figure 6.8: Associative weights for the Q, R, and QR channels and four weights from the QRS channel (to category B) at the end of each block of learning for one run through the type IV category structure.

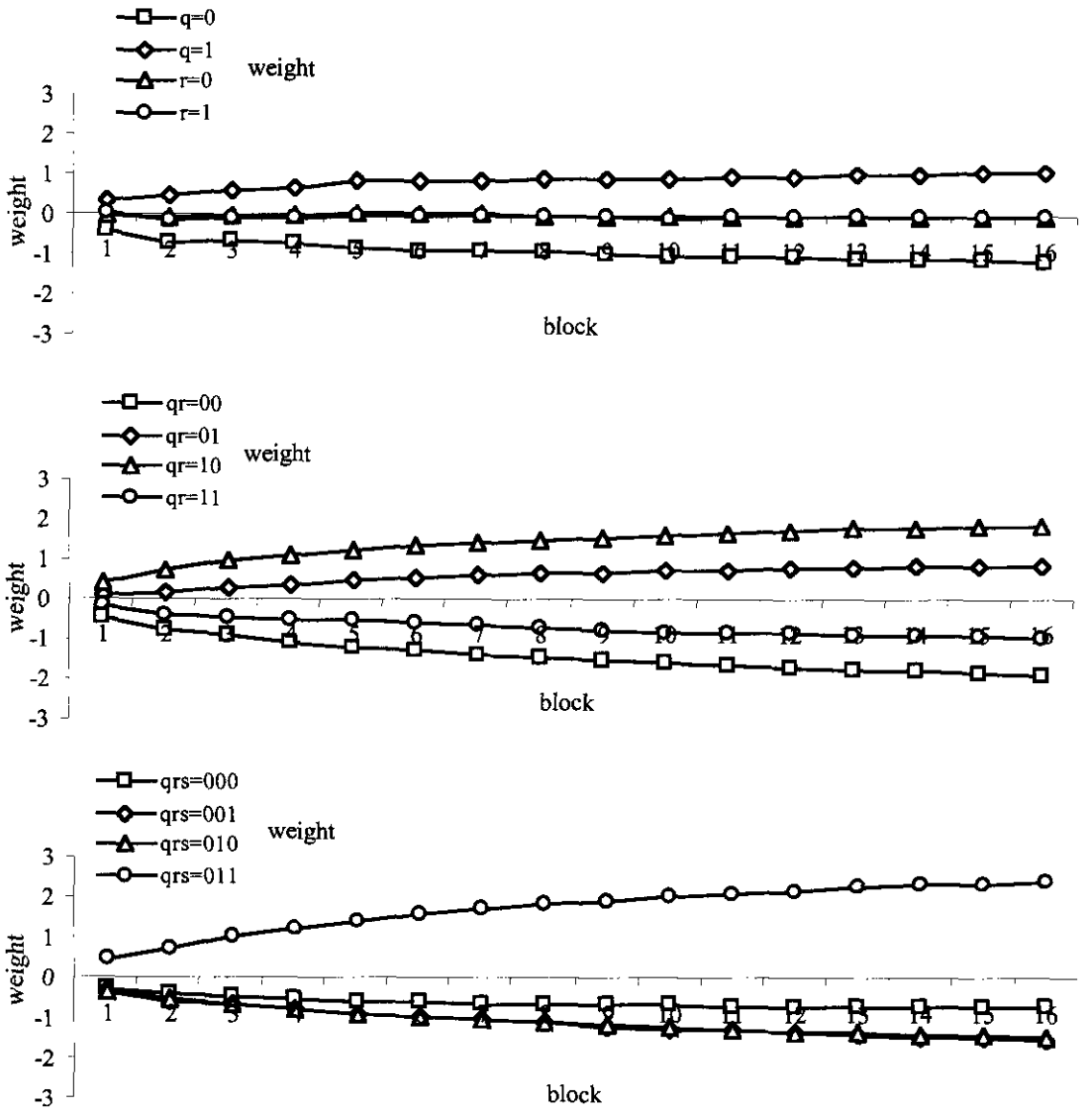

Figure 6.9: Associative weights for the Q, R, and QR channels and four weights from the QRS channel (to category B) at the end of each block of learning for one run through the type V category structure.

Though not shown here, the same pattern is also shown for non-valid sources in the RS channel for the type III task. In this case, weights are in the opposite direction to that suggested by the value of s. Similarly, in the type V task, non-valid QS source weights follow a similar pattern to the non valid QR weights, again having opposite signs to those suggested by the value of q they contain.

Weights do not develop on non-valid two-dimensional sources when both of the dimensions are independently, partially valid. This is shown for the QS channel for the type III task in the third panel of figure 6.7, and all of the two-dimensional channels in the type IV task, one of which is shown in the middle panel of figure 6.8.

Examination of the types III and V category structures (shown in figure 2.2) shows that the conditions under which these non-valid sources occur, given the use of a global teacher signal based learning algorithm, are likely to promote these weights.

The non-valid source is non-valid because on half of the occasions it is present, the stimulus belongs to category A and on the other half, it belongs to category B. In the context of its one, partially valid dimension, half of the times the source occurs the partially valid dimension is 'correct' about the category label, and the other half of the time it is 'wrong'. The non-valid source thus develops weights in the opposite direction to those for its one, partially valid dimension, because error signals will be greater on those occasions when the one partially valid dimension is wrong than when it is right.

Examination of the relevant category structures reveals that the weights developed by these non-valid sources are correct in the context of a peripheral exemplar, and incorrect in the context of central exemplars for the type III structure. For type V they are correct in the context of exception members and incorrect in the context of peripheral members.

Whether this result accurately reflects human performance on the task is, possibly, a question for future research. The effect is somewhat similar to the base-rate effects observed by Gluck and Bower (1988a) described in chapter 3. If one were to test transfer performance on compounds and components of the stimuli following learning of the type III task, say, the model might predict that the participant would be less likely to ascribe qr=01 to category B than q=0. This would be despite the fact that qr=01 is non-valid in isolation, r=1 is non-valid and q=0 is, on 0.75 of trials, a member of category B.

In addition, it is interesting to note that the size of associative weights for the three-dimensional channel, QRS, for tasks III, IV and V correspond to their 'logical status'. In all cases, weights for central members are smaller than weights for peripheral members. In the case of the type V structure, the weights shown in figure 6.9 follow the same pattern, with weights for exception members being the largest of all.

Weights are not shown for the type VI structure but, as might be expected, large weights develop on QRS sources, with all weights in non-valid lower dimensionality channels extremely low.

6.1.4.4. General comments

The approach developed here is somewhat encouraging in that it does appear to be, to some extent, capable of simulating the difficulty levels reported by Shepard *et al.* (1961). It also provides a significant advance on previous configural-cue variants in that it offers a genuinely dimensional attention process, rather than one that requires some form of modular organisation (as with the previous models).

There are, as discussed above, problems with the model. It is, however, a relatively simple model with only four free parameters. Little attempt was made here to optimise these parameters, but it is worth discussing, briefly, their role in the model.

Some reduction in the gap between types III and IV tasks, and the type V task may be achieved by increasing the gain on the decision process. Higher gain values are likely to improve performance on all tasks. The main reduction in the gap between types III and type V is likely to be noticed at the asymptotic end of the learning curve. Increasing learning rate parameters is likely, up to a point, to improve performance on all tasks. As with all learning rate increases, however, it may also make the performance of the model more unstable, particularly with regards to early blocks of training. This is likely to adversely affect early performance on the type II and type VI tasks.

It is also important to note that the associative weight learning rule used in this model will mean that there is some interaction between the associative learning rate and the decision gain. Because the increment to associative weights is a function of the discrepancy between the choice probability and the target, the gain on the decision function will have a role in controlling the size of the increment actually delivered.

As mentioned above, the sampling decision entropy parameter, h, has a role in controlling the probability of configural activation. It exerts a direct control on the diagonal, transitional probabilities $p(i|i)$. Not using the entropy measure at all, or setting the value of h to zero, severely attenuates learning on tasks dependent on configural representations.

Using the entropy parameter tends to promote configural learning. Without it, the late superiority of the type II task is likely to be more pronounced as initial sampling probabilities would favour the tasks with the largest number of valid sources. The lower activation of the configural channels would decrease the rate at which back-propagation of error might 'discover' relevant and non-relevant dimensions. Increasing the parameter tends to enhance configural processing. This does not tend to make much difference to the order of difficulty, but tends to decrease the gap between types III and IV tasks, and the type II task.

The reason for this is that the decision entropy, H(D), at asymptote, is less for the type II structure than for structures III to V. This is because the base of the logarithm used to determine H(D) is three, such that entropy is maximal when all three probabilities are equal.

This means that for the types III and IV tasks, higher values of h tend to enhance the activations of all of their configural channels. For the type II task there is an initial problem with the enhancement of learning in the QRS channel during early blocks. This tends to slow down the operation of the back-propagation process as the three-dimensional channel provides positive signals to all dimensions. This reduces the rate at which the sampling probability for the irrelevant dimension is reduced. This, in turn, decreases the rate at which the two-dimensional channel increases its activation over the basic level reached for equal initial sampling probabilities. The result is a reduction in the advantage of the type II over the types III and IV tasks.

This effect is representative of a major obstacle to the potential generalisability of the model to other learning tasks. If one had a task with many dimensions of which only two were relevant, as a compound only, the discovery of this relevance would drastically reduce the sampling probabilities for all but two of the dimensions. The entropy of the decision would, being to a much higher base, be comparatively low. This would increase the probability of consecutive sampling of the component dimensions of the relevant configuration and thus may considerably reduce configural activation.

Ideally, once a dimension has been learnt to be irrelevant, its influence over the utilisation of dimensions which *are* relevant should be attenuated. Although the influence of irrelevant dimensions is indirect in this model, it is, potentially, problematic.

The problem of 'disembodied' probabilities is another difficulty for this model. The initial sampling probability vector, **p**, refers to the probabilities of sampling, without any explanation as to the system which gives rise to the probabilities in the first place.

Probabilities, when normalised in this way, refer to relationships between events; q is more likely than r, for example. The model implies that some learning process makes q more likely to occur than r but does not necessarily suggest how this may be brought about. The model may be developed by replacing the probabilities with some form of weight that is altered by the back-propagation process. In this way one might suggest that the normalised weights refer to the initial sampling probabilities for the dimensions involved. If the weights began with non-zero values, then it would be possible to incorporate new dimensions into the model and attempt to model experiments where the number of stimulus dimensions, or components, was variable.

The following model addresses, directly, both of these issues. It attempts to represent the dimensional attention aspect of the learning process in terms of a set of *relationships* between dimensions, rather than in terms of characteristics of the dimensions themselves. The difference is that the adaptive components with respect to the sampling process are the transition probabilities, rather than the initial probabilities.

6.2. Alteration of transition matrix by back-propagation of error: the Adaptive Transition Matrix (ATM) model

The second model attempting to implement a form of dimensional attention uses back-propagated error to specifically affect the transition probability matrix **P**. For this model, source node activation is still dependent on the sampling probabilities for individual dimensions and sequences of consecutively sampled dimensions. In this case the initial probabilities, **p**, remain fixed at one over the number of dimensions (i.e. 1/3). The values that change from trial-to-trial are the transition probabilities **P**.

6.2.1. Determination of transition probabilities

For this model, the sampling process may be represented in terms of a fully interconnected network of three nodes, one for each dimension. This arrangement is shown

in figure 6.10. The probability of a node being sampled at time t+1, or $p(d)_{t+1}$ is given by the following;

$$p(d)_{t+1} = \frac{a_{d(t)}}{\sum_{d} a_{d(t)}} \qquad (6.7).$$

The activation of a dimension node, D, a_D is dependent on the input from all of the nodes, d, in the sampling network,

$$a_D = \sum_{d} \tau_d \alpha_{dD} \qquad (6.8).$$

The term τ_d is the transmission 'rate' of the dimensional node d. This term takes on a value of one, if the dimension is the dimension which is currently sampled, and zero otherwise. The effective connection strength between dimension d and dimension D is given by α_{dD}. This connection strength depends on the value of an adaptive unbounded positive or negative weight θ. The effective connection strength between dimensions i and j (where i may be the same dimension as j) is maintained in the unit range as follows;

$$\alpha_{ij} = \frac{1}{1+e^{-\theta_{ij}}} \qquad (6.9).$$

Because only one dimensional node may be sampled and thus transmit at any one time, the conditional probabilities in the transition matrix **P** are thus determined just using the effective connection strengths. The conditional probability $p(j_{t+1}|\, i_t)$ is evaluated by the following;

$$p(j_{t+1}\,|\,i_t) = \frac{\alpha_{ij}}{\sum_{k} \alpha_{ik}} \qquad (6.10),$$

where k are all of the dimensions including i and j.

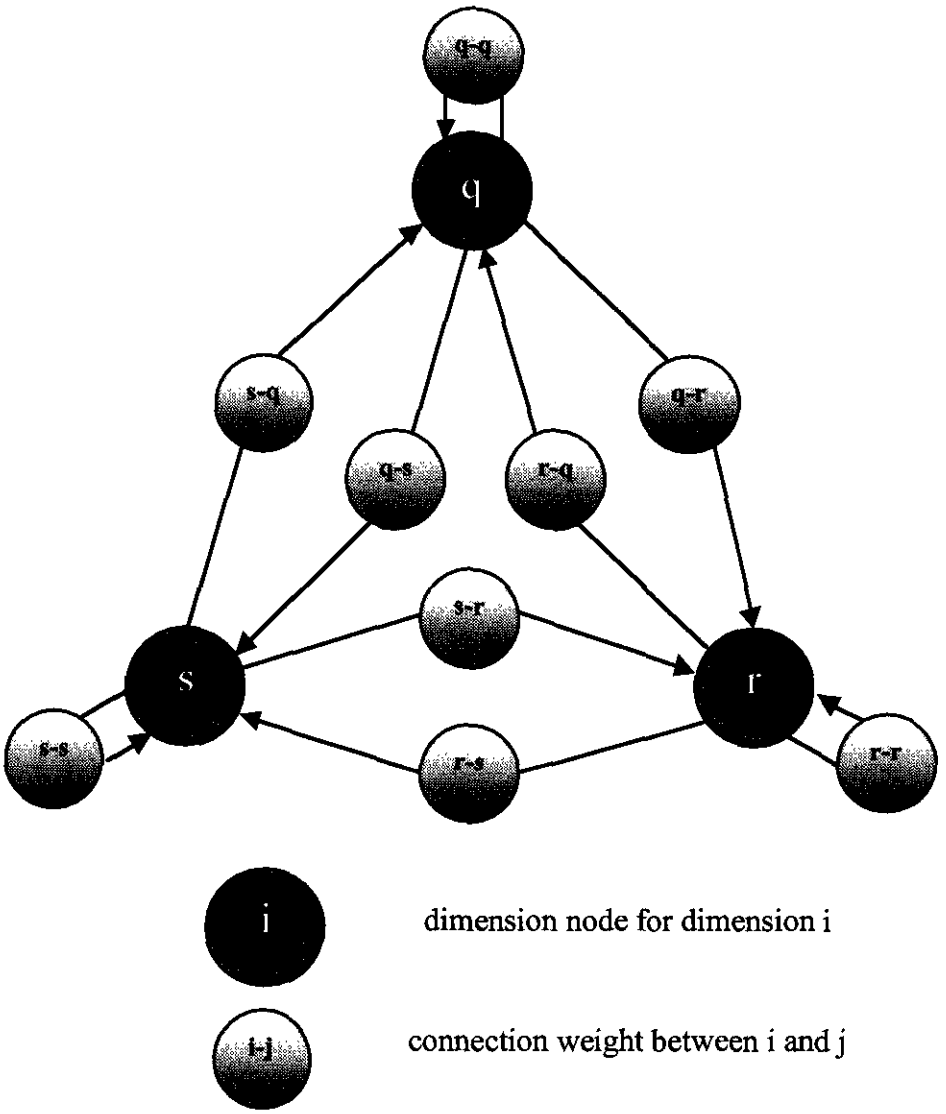

Figure 6.10: Representation of the sampling system as a system of three dimensional nodes interconnected via adaptive weights.

6.2.2. Feed-forward activation functions

This model employs the same activation functions as the ASP model. The choice probability is determined using the summed channel output, as in equation 6.1. The channel output is determined by the maximum weighted contribution and channel activation probabilities as in equations 6.2 and 6.3.

The channel activation probabilities are calculated using similar methods to those given for the previous model. An important difference, in this case, is that there is no 'global' uncertainty term H(D). The probability of consecutive sampling of the same

dimension twice is controlled by the transition probability $p(i_{t+}|\, i_t)$ determined, along with the other transition probabilities, using equation 6.10.

This model makes use of the specific probabilities of particular sequences of samples in determining the effects of back-propagated error signals on the transition weights. As such it is worth defining the channel activation probabilities in terms of these measures, in order to introduce the notation required in the next section.

The probability of a channel being active during the trial is given, as for the ASP model and for the dimensional attention model in chapter 4, by the sum of the probabilities of the different sequences of samples which may activate it. As described above, a channel is activated by consecutive sampling of its component dimensions. For example, equation 4.26 gives the activation probability of the QRS channel as being the sum of the sequence probabilities; p(qrs), p(qsr), p(rqs), p(rsq), p(sqr), and p(srq).

In this case a channel, c, has a set of U dimensions $u_c = (d_1, d_2 \ldots d_U)$ and it is activated by any one of a set of sequences of samples of those dimensions. A sequence 'member' of c is described by $v_c = \{d_t, d_{t+1} \ldots d_{t+(U-1)}\}$. Each element of the sequence is a different element of u_c. Where U for a particular channel c is one, i.e. it is a one-dimensional channel, the probability of the channel being active, p(c), is just given by p'(d). This is the estimated probability of the dimension being sampled following Z steps of the sampling process, determined as with the previous model and the model in section 4.3.3.2. For a multidimensional channel, the probability of channel activation during a trial, p(c), is given by the following;

$$p(c) = \sum_{m=1}^{U!} p(v_c)_m \qquad (6.11).$$

The probability of the sequence occurring during the trial being given by;

$$p(v_c) = p'(d_t) \prod_{t=1}^{U-1} p(d_{t+1} \,|\, d_t) \qquad (6.12).$$

6.2.3. Updating transition and associative weights

At the end of each trial, associative weights and the weights which determine the transition probabilities in the sampling process are updated. Associative weights are altered according to the same rule as for the previous model, given in equation 6.5. The functions

governing the alteration of transition weights are somewhat more complicated than those used in the previous model.

The basic heuristic used was to increase a transition weight between its origin and destination, if either that transition is involved in activating valid representations, or the destination is generally more involved in more valid relationships that the origin. The measure of the general contribution of a particular dimension to the decision is based on the back-propagation of error from channels dependent on that dimension, via transitions which *end* with that dimension. The signal back-propagated to a dimension, d, or b_d is *determined by the following function,*

$$b_d = \sum_{\substack{c \\ d \in u_c}} (\delta_l - p(l)) o_{cl} \frac{\sum_{\substack{m \\ d=d_{i+(U-1)}}} p(v_c)_m}{p(c)} \qquad (6.13).$$

This is basically the error back-propagated through channels of which d is a component, multiplied by the fraction of that channel's activity which may be attributable to sequences ending in d.

The signals for individual weights differ, depending on what kind of weight they are. The update signal for the weight between a dimension and itself is different to the update between two different dimensions. The change signal for a connection between a dimension and itself is a function of the amount of signal back-propagated from the channel that is dependent only on the dimension itself, and the difference between that signal and the signal from multidimensional, dependent channels,

$$\Delta\theta_{ii} = \left((\delta_l - p(l)) o_{il}\right) + \left[p(ii) \left(\left((\delta_l - p(l)) o_{il}\right) - \sum_{\substack{c \\ i \in u_c}}^{U>1} (\delta_l - p(l)) o_{cl} \frac{\sum_{\substack{m \\ i=d_{i+(U-1)}}} p(v_c)_m}{p(c)} \right) \right] \qquad (6.14).$$

The first part of the equation is the signal from the channel for dimension i. The second part of the equation is the difference between this signal and the signal that is transmitted back via dependent channels with more than one dimension, via sequences which end with the dimension i itself. This second part is multiplied by the probability that consecutive sampling of i occurs during the trial. It is basically a measure of what is gained by

consecutive sampling of the same dimension, minus a measure of what will be lost by that sequence, all multiplied by the probability that the consecutive sampling occurs at all.

The update signal for a connection between two different dimensions, i and j, is a function of the differences in back-propagated signals between the destination and origin, and the amount of back-propagated signal which depends on the transition itself.

$$\Delta\theta_{ij} = \left(\left(b_j - b_i\right)\alpha_{ij}\right) + \left[\sum_{\substack{c \\ i,j\in u_c}} \left(\delta_l - p(l)\right) o_{cl} \frac{\sum_m p(v_c)_m}{p(c)}^{i=d_{U-1}, j=d_U}\right] \qquad (6.15)$$

The first part of the equation is the difference between back-propagated signals to the destination and origin dimensions. This is weighted by the current value of the effective weight between the two, α_{ij}, to reflect that once it has been learnt that a destination dimension is irrelevant, the difference between the signals to the two dimensions is of reduced importance. The second part of the equation consists of the sum of signals back-propagated via channels which are dependent on the transition.

The value of the weight on the next trial is then computed by the following,

$$\theta_{ij(t+1)} = \theta_{ij(t)} + \left(\Delta\theta_{ij(t)} - \sum_{k\neq j} \Delta\theta_{ik(t)}\right)\lambda_\theta \qquad (6.16).$$

This means that the overall change to the weight θ_{ij} is determined by subtracting the increment signals to all other transition weights leading away from dimension i, from the increment signal for θ_{ij}. The parameter λ_θ is a learning rate constant.

As can be seen, the learning functions for these transition weights involve a high degree of competition. For the connections between a dimension and itself, the update signal shown in equation 6.14 involves the signals to the dimension in isolation being compared with the signals to transitions to other dimensions. Where the dimension is more valid in conjunction with other dimensions than it is by itself, this recurrent connection is likely to receive a negative signal.

Equation 6.16 introduces further competition in that updates to the transition weights are dependent on whether that transition is *uniquely* relevant or not. As with the RMAW model, and similar to the scheme proposed by Mackintosh (1975), only the weight with the highest update signal will receive a positive change.

6.2.4. Results from the experimental simulation and discussion

The experimental simulation was carried out in the same way as for the previous models. The model was run through the sixteen training blocks of each experiment twenty times, with a different randomised order of stimulus presentation for each run.

6.2.4.1. Overall performance of the model

The parameters used for the results which follow were λ_w=1, g=2.5, and λ_θ=4. In addition the transition weights for connections between a dimension and itself θ_{ii} were initialised at –0.5, with all other transition weights initialised at zero. This was done in the hope that the problems with the early performance on the type VI, suffered by the previous model, would be, to some extent, rectified.

The ASP model has a form of 'built-in' early enhancement for configural activation, in that the equal initial sampling probabilities results in a maximal value of H(D) for early trials. This model lacks such a provision and so, without this pre-setting of 'recurrent' transition weights to lower values, severely attenuated early learning in the types II and VI structures could be expected.

Figure 6.11 shows the average probability of the correct response being produced by the model, per block, averaged across the twenty runs through each category structure. Again, the model exhibits the order of difficulty observed by Shepard *et al* (1961) and Nosofsky *et al* (1994). Performance on the type II task is, for this model, considerably better than that shown for structures III to V. First-block performance on this task, while not superior to that on tasks III and IV, is certainly not clearly worse.

Performance on structures III to V is interesting in terms of the fact that while the overall probabilities of correct responding across the sixteen blocks are fairly close for the three tasks, the learning curves are somewhat different in nature. The average probabilities of correct response per block across all sixteen blocks are 0.906, 0.891, and 0.875 for types III, IV, and V respectively.

The curves in figure 6.11 indicate that the errors tend to be distributed across the learning curves in different ways. Type IV seems to suffer from poor performance early on in learning, but then its performance increases to the highest asymptotic levels of the three. Type V has superior early performance but its performance towards the end of the experiment is noticeably worse than that on types III and IV. The performance on the type

III structure begins in a similar fashion to that for the type V task, but higher levels of performance are achieved towards the end of the sixteen blocks.

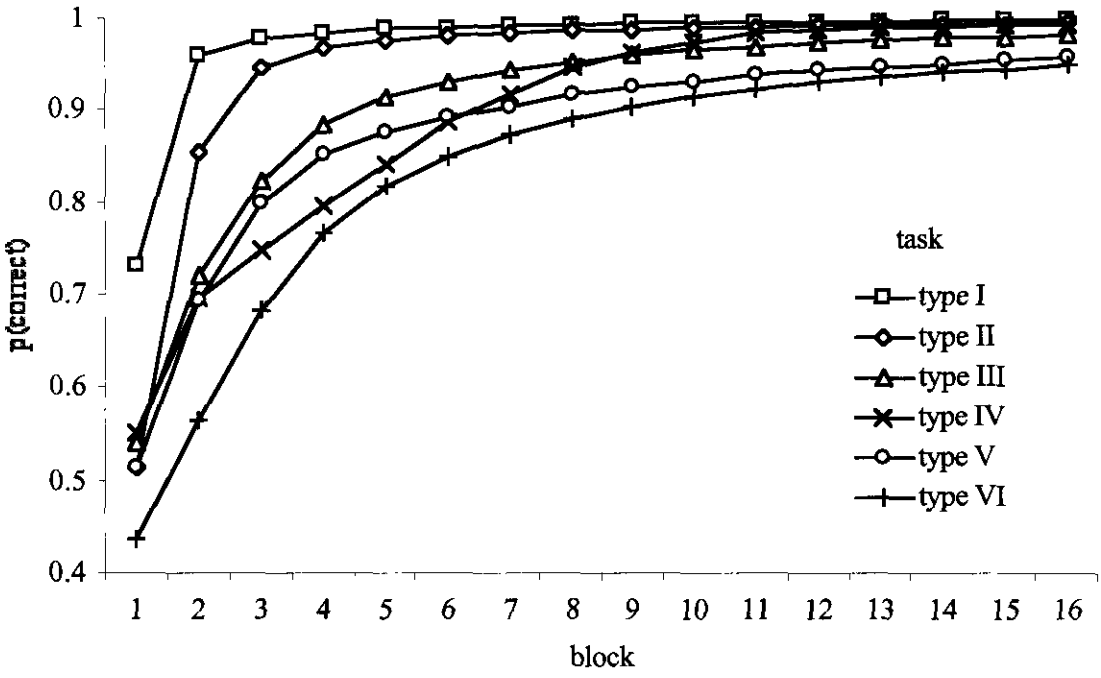

Figure 6.11: Mean p(correct) per block averaged across the twenty runs through each task for the ATM model.

The reasons for this particular pattern will be discussed below. Analysis of the transition weights and channel activation probabilities reveals that the model tends to behave in distinctly different ways for each of these three tasks.

Performance on the type VI task is still somewhat poor in comparison with the human data for the early trials. Despite the fact that its performance towards the end of the task is only marginally worse than that on the type V structure, its overall performance is clearly worse than that shown for the rest of the category structures.

Figure 6.12 shows how well the model learns patterns with different logical status in the types III to V tasks. As with all of the previous models, central members are easier to learn than peripheral ones which are, in the case of the type V task, easier than exception members.

6.2.4.2. Channel activation probabilities and transition weights

Figure 6.13 shows the average probabilities of channel activation, averaged across each block, and across the twenty runs through each category structure. As the figure shows, for the type I task the activation for the valid Q channel rapidly increases towards unity with that of all other channels heading towards zero.

The graph for the type II structure shows similarly rapid 'learning' of the relevant dimensions of the task. Unlike the previous model, activation of the valid QS channel on this structure approaches one.

Both of these patterns can be understood in terms of the transition weights that develop between dimensions. Figure 6.14 shows the average, normalised, output from each dimension. This is the conditional probability of the destination dimension being sampled at time t+1, given that the origin dimension has been sampled at time t.

As the output is the weight θ_{ij}, passed through a logistic, as given in equation 6.9, the indication is that, for the type I structure, very high weights develop between all dimensions and q, with the weight from q to itself also being very high. All other weights are strongly negative.

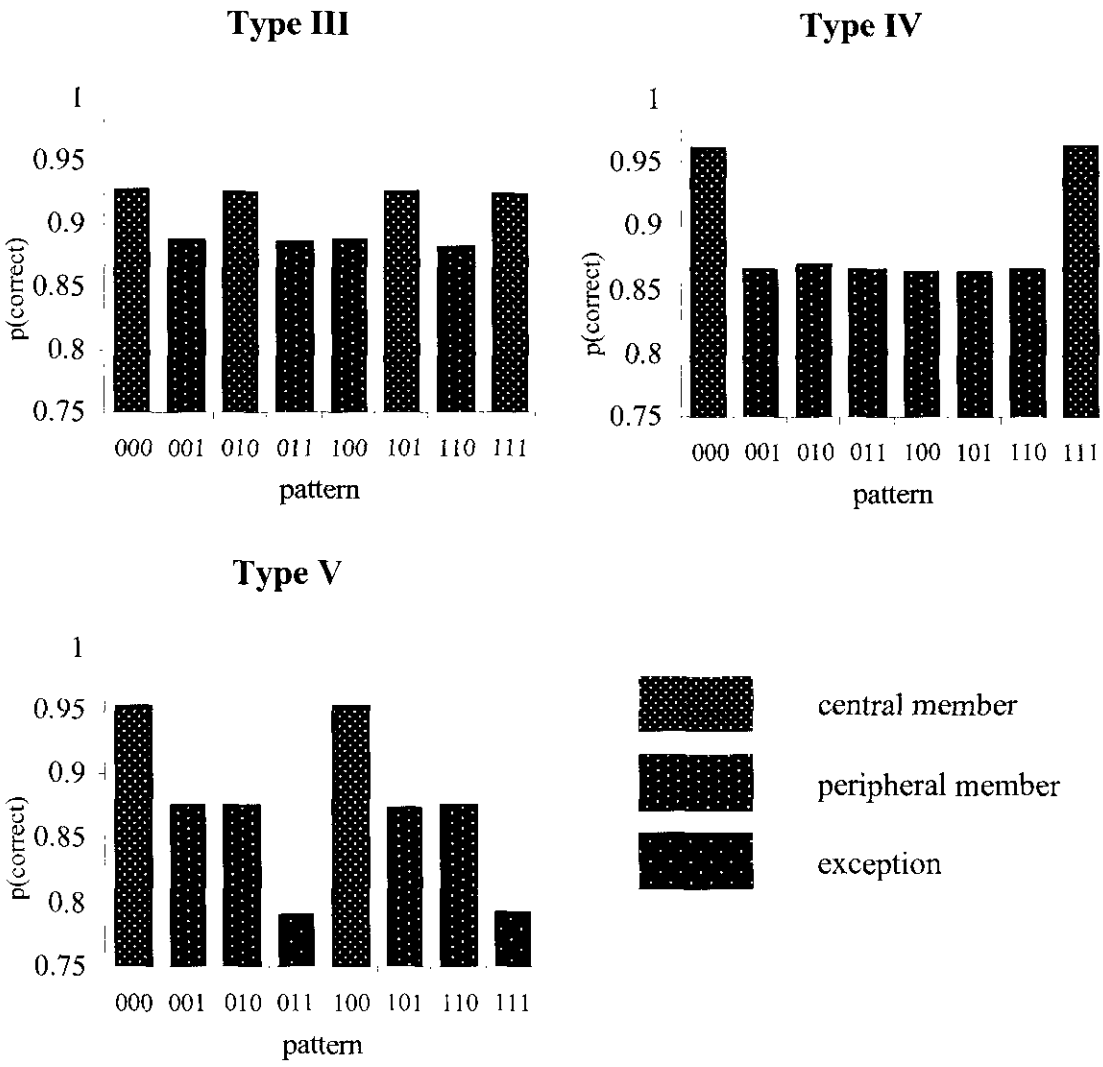

Figure 6.12: Performance of the model on individual patterns for category structures III, IV, and V. Performance is indexed in terms of the average probability of correct responding across the entire 16 blocks of the simulation. Figure 2.2 shows the relationships between these three types of category member in each category structure.

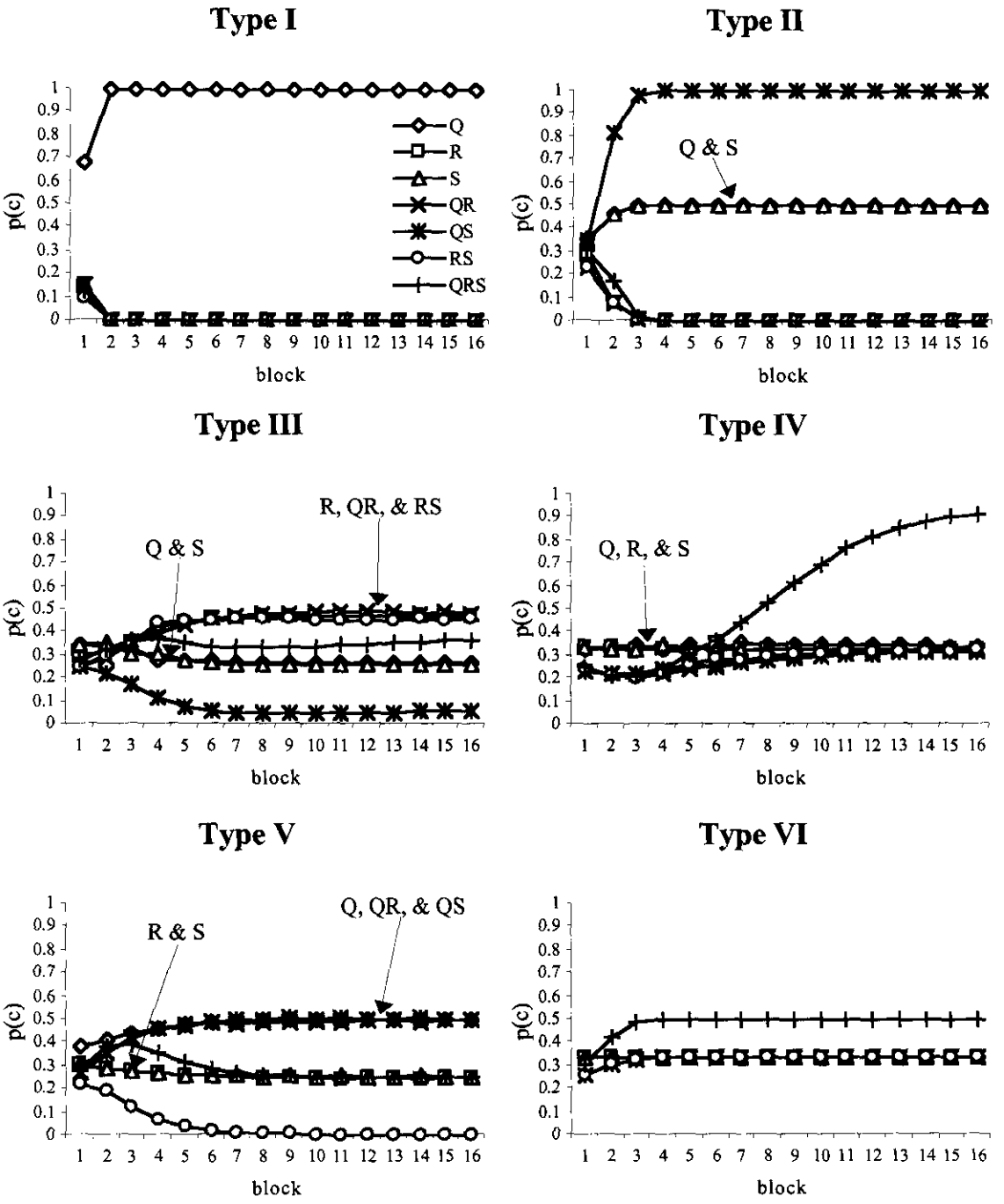

Figure 6.13: Average probabilities of channel activations, p(c), per block. Key shown for the type I task applies to all graphs.

For the type II task, the high probability of activation in the QS channel can be understood in terms of extremely high weights between the q and s dimensions. Unlike the previous model, this model can control the probability that it will sample the same dimension twice in a row in terms of its weight to itself. In this case the weights for q to q and s to s are strongly negative. This results in a pattern of sampling that oscillates between q and s, resulting in the high probability of activation shown in figure 6.13.

For the type VI task all weights between dimensions were positive, with all weights between a dimension and itself being strongly negative. As these weights become increasingly negative, the probability of the QRS channel being active, shown in figure 6.13, increases towards 0.5. As described in section 6.1.4.4, this is the probability of the QRS channel being activated when the probability of consecutive sampling of the same dimension is zero.

The channel activations and transition probabilities for structures III to V show different characteristics from one another, as suggested by the curves in figure 6.11. For the type III structure the pattern observed is one of high activation probabilities for the *non-valid* R channel and the partially valid QR and RS channels. With respect to the single-dimensional channels, asymptotic activation probabilities are approximately 0.25, 0.5, and 0.25 for Q, R, and S respectively. The normalised effective transition weights, shown in figure 6.14 for this structure, suggest that the r dimension will be sampled on approximately every other step of the process. As described for the ASP model, and also for the RMAW model, in a situation where learning is somewhat competitive, the QS channel is likely to develop at a lower rate than RS and QR channels in this structure.

It is important to note that the model does retain some level of activation in the QS channel and, from figure 6.14, there is still a small chance that q-s and s-q transitions will occur. When comparing the performance on the type III with performance on the type V, this would appear to be important.

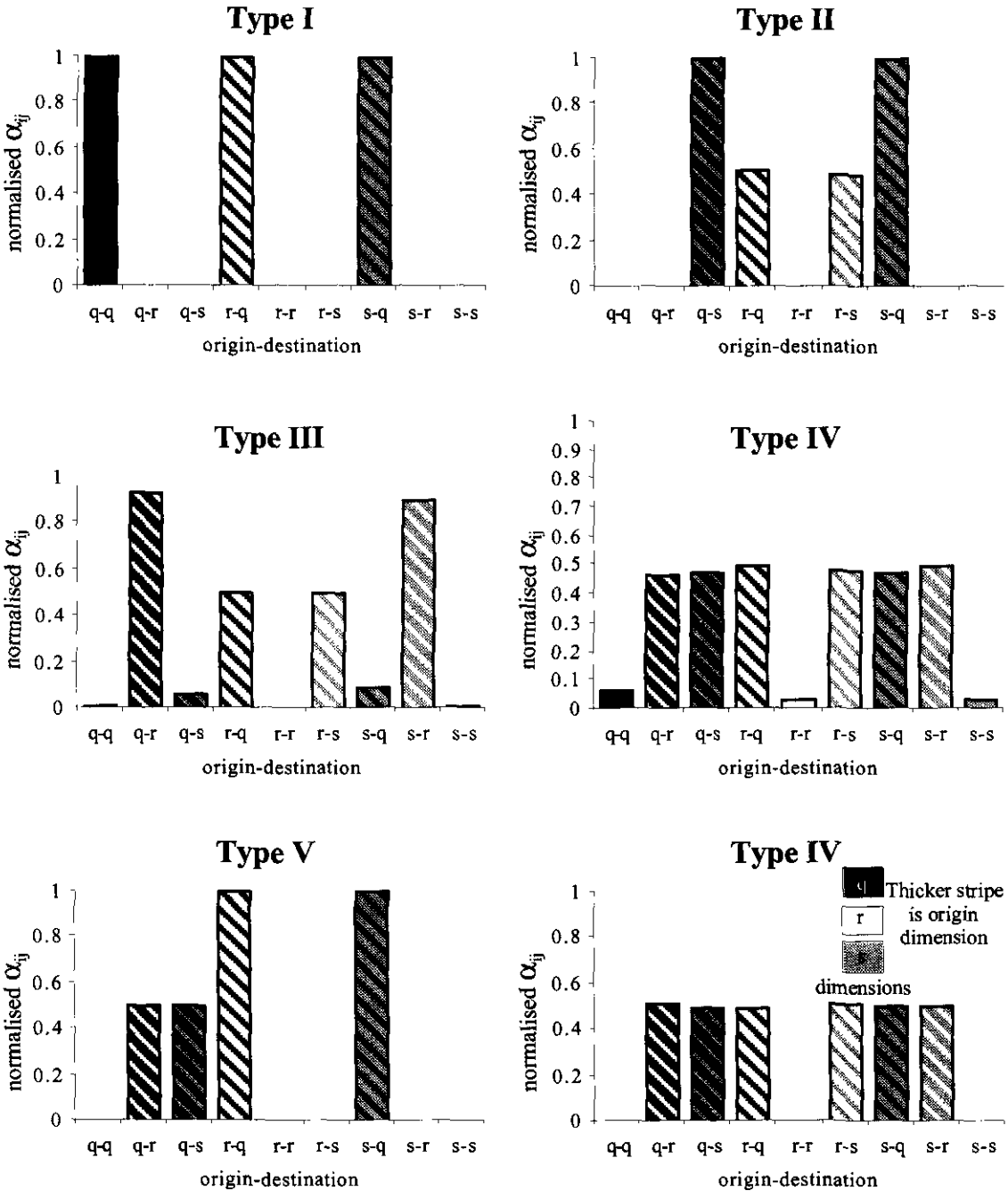

Figure 6.14: Average effective transition weights (α_{ij}), averaged across the last four blocks of training, normalised across each origin dimension, e.g. q-q + q-r + q-s = 1. This is effectively the conditional probability of the destination being sampled at time t+1, given that the origin has been sampled at time t. Key shown for type VI applies to all charts.

On the type V structure, the model tends to sample the single valid q dimension in preference to the non-valid r and s dimensions. As figure 6.14 shows this tends to result in a similar pattern of sampling where, in this case, q is likely to be sampled on every other step of the process. Unlike the type III structure, however, the members of the categories on the type V structure cannot all be correctly classified in terms of the valid sources in the two partially valid two-dimensional channels.

Because the reduction in the activation of the non-valid RS channel is more or less complete in this task, the probability of the QRS channel being activated is attenuated. This channel is most useful, in this task, for exception members and the attenuation of its activation severely degrades performance on these members.

Activation of the QRS channel is dependent on the consecutive sampling of the three dimensions. The pattern of transition weights developed in the type V task allow this to occur at a maximum rate given by the sequence s-q-**r**-q-**s**-q-**r**-q-**s**, where bold type indicate the activation of the QRS channel. This would allow the channel to be activated approximately half of the time assuming that q was sampled on every other step of the process. Unfortunately, following sampling of q the chance that the next dimension sampled will be the one that preceded q is equal to the chance of sampling the remaining dimension. This reduced the probability of QRS activation to about 0.25, as shown in figure 6.13.

The situation is slightly different for the type III structure. While the transitions are predominantly between r and q and r and s, there remains a significant chance that a transition between q and s will occur. This enhances the probability of QRS activation by increasing the chance that consecutive samples will result in consecutive activations of the QRS channel via, say, the sequence q-**r**-**s**-**q**. While the s-q transition is less likely than other transitions, when it does occur, it tends to result in consecutive activations of the QRS channel. Note that if the above sequence does occur, the most probable next sample is dimension r, such that the QRS channel would be activated on three consecutive steps of the sampling process.

The pattern of activation and transition probabilities for the type IV structure shows an entirely different tendency. In this task, all of the single dimensions are equally (partially) valid, and all of the two-dimensional channels contain the same number of valid

sources. For the central category members, all of the partially valid channels are fully valid. For the peripheral members two of the single dimensional channels are valid, and the two dimensional channel, of which the two valid dimensions are components, is also valid.

Figure 6.13 shows that the model retains equal channel activation probabilities for each of the three single-dimensional channels. The configural channels begin with a somewhat lower activation probability, but then increase. This increase is most marked for the three-dimensional channel, which increases towards a probability of about 0.85 by the end of the experiment.

The normalised effective transition weights shown in figure 6.14 indicate that, on average, the model has developed equal transition probabilities for each transition to a different dimension with the probability of consecutive sampling of the same dimension low, but not as low as that for the type VI. In fact, apart from these consecutive sampling probabilities, the average matrix appears to be similar to that for the type VI.

As will be noticed from figure 6.13, the probability of activation for the QRS channel in the type VI task is much lower than that for the type IV task. The reason for this is that on the type IV task the model develops transition weights that support *cyclical* patterns of sampling. Closer inspection of the normalised effective transition weights reveals that individual networks either develop a 'clockwise' (q-r-s direction) or 'anticlockwise' (s-r-q direction) pattern of weights. There were 10 of each type. Figure 6.15 shows the averaged normalised effective transition weights for these two groups.

The reason why cyclical sampling patterns develop for the model on the type IV structure is somewhat difficult to elucidate. The learning algorithms for the transition weights suggests that once cyclical sampling occurs, it is likely to persist, particularly in structures where all dimensions are equally valid. Because the signal back-propagated to a transition weight is a function of the frequency of that transition, once 'travelling' in one direction is more common that the other, the dominant direction is likely to receive more error signals.

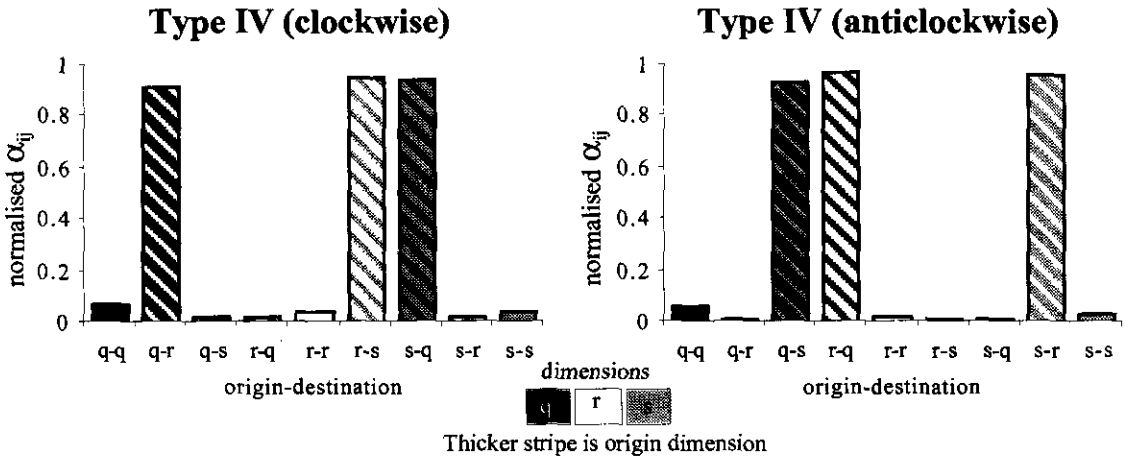

Figure 6.15: Average normalised effective transition weights (α_{ij}), averaged across the last four blocks of training for the type IV task, separated into those runs which resulted in 'clockwise' and 'anticlockwise' transition weight matrices (10 of each type).

Furthermore, changes to a transition weight, due to differences between the back-propagated signal to the origin and destination of the transition, are gated by the size of the transition weight between them (as in equation 6.15). This means that once a transition weight has been reduced, signals due to this source are unlikely to alter it a great deal.

The enhanced contribution of the three-dimensional channel which results from the cyclical sampling pattern is also likely to enhance the process. The only time that negative signals to transition weights between dimensions will occur in this task, is when peripheral category members are presented and the back-propagated signal to a destination dimension will be less than that transmitted to the origin. The signal back-propagated via the three-dimensional channel to this transition weight will always be positive. Furthermore, the magnitude of this contribution will be greater for peripheral stimuli as the three-dimensional associative weights for these members are greater. When central stimuli are presented all transitions and dimensions are equally valid, therefore enhancement of transitions will be greatest for those transitions which occur most frequently.

Of all of the category structures the transition weights for the early blocks of training on the type IV are the most unstable. The learning rate parameter was set to quite a high value for these results (4). This, coupled with the competitive algorithm, the equal

validity of all dimensions and transitions, and the fact that not all of the dimensions and transitions are valid on all trials, is likely to account for this instability. If cyclic patterns of sampling 'emerge' from this initial instability at some point they may, due to the reasons described above, be 'self-promoting' or act as stable attractors for the weights.

6.2.4.3 Associative weights

The associative weights that developed for the model are not shown here as they are substantially similar to those that developed for the ASP model. The weights for redundant relevant channels in the types I and II task were slightly smaller, this being accounted for by the more rapid and complete development of sampling patterns which preclude their activation.

Development of weights on the ostensibly non-valid two-dimensional sources for the types III and V tasks followed a similar pattern to that shown for the ASP model. In this case the development was slightly less in the type III task and slightly greater in the type V task. Despite the high level of activation in the three-dimensional channel, the size of QRS associative weights in the type IV task was no greater than that for the types III and V tasks. This may be accounted for by the fact that by the time the three-dimensional channel becomes dominant, substantial weights on the lower dimensionality channels have already developed.

6.2.4.4. General comments

The ATM model represents another way of implementing a form of dimensional attention using the configural-cue form of representation. Like the ASP model it achieves a qualitative fit to the data reported by Nosofsky *et al.* (1994) shown in figure 2.3, which is superior to that of the configural-cue variants reported by these authors (*ibid.*).

The model represents something of a conceptual improvement over the ASP approach. As discussed in section 6.1.4.4, it seems likely that without further alterations it may be difficult to apply the ASP model to situations in which stimulus dimensionality varied.

The ATM approach represents the selective attention process as dependent upon learnt relationships between dimensions, in the context of the task. New dimensions may be introduced with the assumption that the transition weights connecting them to each other have an initial value of zero. How this representation of new components or

dimensions might affect subsequent learning is, to some extent dependent on the way in which one models the operation of the 'attention' learning process.

The next model attempts to clarify this issue by representing the process in a much more specific way. As will be seen, the way in which this may be achieved is by assuming that the process operates in a similar way to that used in the rapid attention shift models, described in section 3.3.5.

6.3. Rapid attention shifts: the Rapidly Adaptive Transition Matrix (RATM) model

The final model presented in this chapter is an attempt to address some of the issues raised by the previous two. One issue concerns the direct representation of source node activation in terms of the probability of channel activation. With the ATM and ASP models there is no model offered as to how this probability is 'converted' into node activation. It may be suggested that this is an undesirable characteristic for a connectionist model.

The use of channel activation probabilities also extends to interpretations of the various learning rules used. The sampling process in the previous two models is represented as a sequential process. Probability or transition weight update signals were calculated in terms of the probabilities of informative transitions, or samples, being made in relation to the feedback delivered for a particular stimulus. The above models thus attempted to approximate a process which would require time, using average measures of the likelihood of each transition occurring.

Representing learning for the sampling process more explicitly would necessitate a number of iterations of the sampling process to occur, after the feedback had been presented, to allow the valid relationships to be learnt. This means that the sampling process for a model incorporating an explicit representation of sampling would not be representable in terms of a 'within-trial' Markov process. With the explicit representation, the probability of each dimension or transition may be changing on each step of the trial.

Modelling the alteration of sampling probabilities using this sequential process is reminiscent of the rapid attention shift models, such as ADIT and EXIT (Kruschke, 1996a, and in press a). As discussed in section 3.3.5, these models alter 'attention' weights for individual representations, or dimensions (in the case of RASHNL (Kruschke & Johansen, 1999)), in such a way as to minimise the discrepancy between the network's output vector and the feedback vector. After these weights have been altered across a number of iterations, the associative weights are changed.

The ATM model, which makes use of sampling probabilities to control the activation of representations, requires that activation be evaluated according to the characteristics of the matrix of transition weights. These weights are modified by learning.

Representing the channel activation as something which is dependent on the sampling process also requires that there be a number of iterations of the sampling process prior to the decision being made. This is necessary to allow the channel activations to 'develop' to reflect the characteristics of the transition matrix.

To implement learning of the matrix of transition weights using a more explicit representation of the sampling process, as stated above, requires some measurement of the effectiveness of particular transitions. This will not necessarily affect the operation of the back-propagation component of the learning process used in the previous model, although in this case the channel activation probabilities must be replaced with some other measure of channel activation.

The simplest way to index the effectiveness of individual transitions would be to measure error before and after they have taken place. If a particular transition results in a decrease in the overall error, then that transition weight may receive a positive update signal. Conversely, if a transition leads to an increase in error, it would receive a negative signal.

The resultant model is, therefore, somewhat more complicated than the previous two. This extra complexity is made necessary by taking into account the development of channel activations across a trial. It is, perhaps, more 'connectionist' than the previous two models, as all of the node activations are the result of explicitly represented transmission relationships, rather than just probabilities.

The only probabilities which remain in the model are those which determine which dimension will be sampled on a given step of the trial. The model is deterministic, however, in that a random number in the unit range was generated at each step of the sampling process, its value being compared to the probabilities generated by the previous step to decide which dimension would be sampled on the current step.

Figure 6.16 illustrates the architecture of the model. As can be seen, the model has recurrent aspects that were not present in the previous two models. The transition weight update rule, which will be described below, relies on both back-propagation of error and

direct feedback from the decision process. In addition, as will be detailed below, the 'state' of the decision process directly affects the characteristics of the matrix of transition weights.

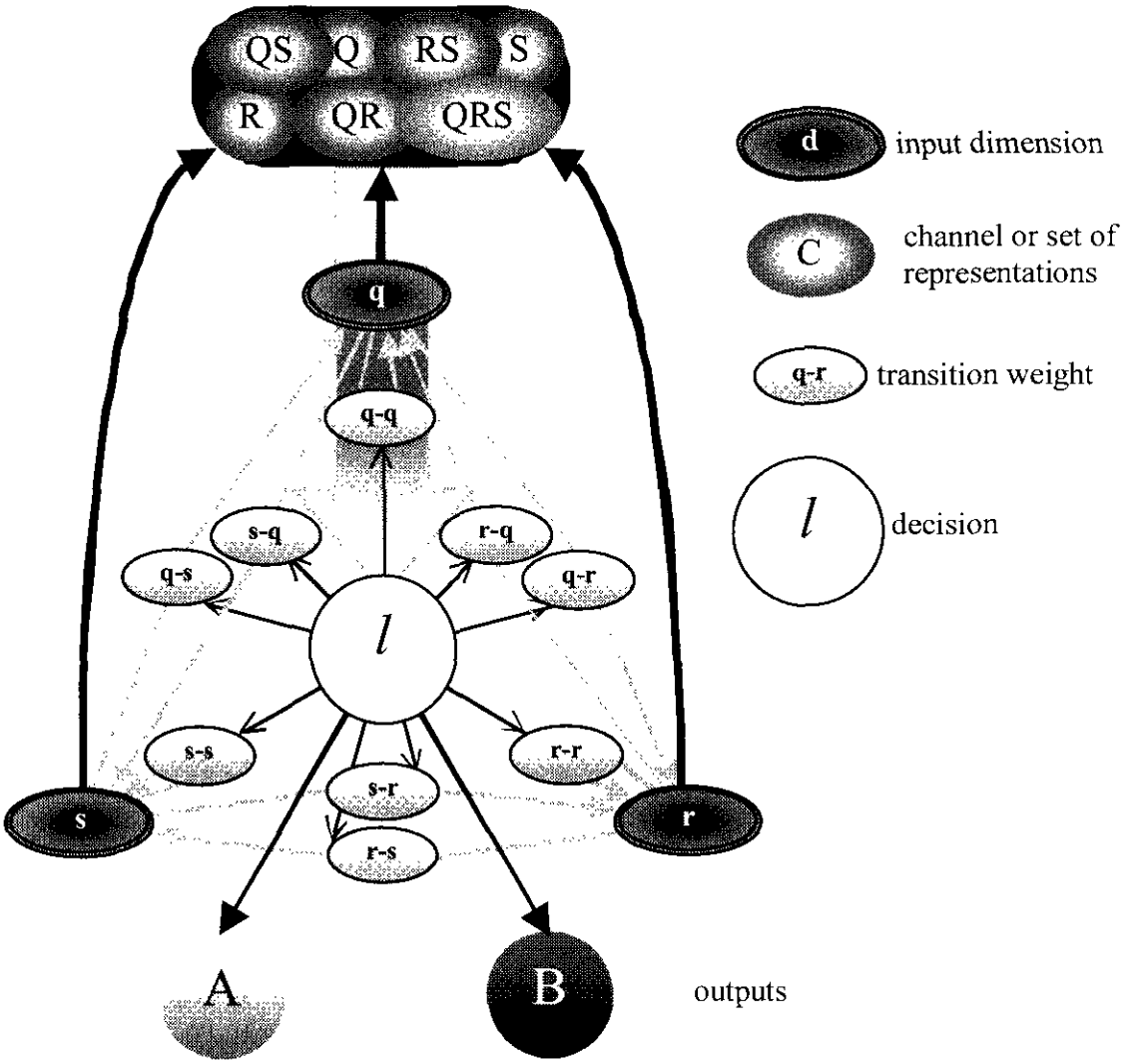

Figure 6.16: Rapidly adaptive transition matrix (RATM) model showing effective connections between decision process and transition weights. There is full connectivity between all source nodes in channels and the decision process, represented by block arrow.

6.3.1. Representation of the trial and the determination of sampling probabilities

For this model, each trial consists of a number of steps. For the results presented each trial, indexed **t**, consists of 75 steps, indexed t. The trial is split into a pre-response phase and a post-response phase. The pre-response phase lasts for 49 steps, with the decision made on the fiftieth step and feedback presented from this point onwards until the end of the trial.

The dimensional sampling probabilities are determined using the same method as in the previous model, using equations 6.8 and 6.9. Unlike the ATM model, the probabilities generated here are used on each step to determine the actual dimension sampled, and thus the value of τ_d for each dimension.

The effective connection strength in this model is determined, from the unbounded weight θ, in a somewhat different fashion. In this case the extent to which the learned connection strength is implemented at a particular step is modulated by a variable relating to the number of steps which have occurred since the trial began. It is also adjusted by a parameter relating to the ongoing uncertainty (pre-response), or error (post-response), regarding the decision. The role of this uncertainty or error-based parameter is different, depending on whether the connection is between a dimension and itself, or between two different dimensions. For a recurrent connection, the effective connection strength at step t, $\alpha_{ii(t)}$, is given by the following;

$$\alpha_{ii(t)} = \frac{1}{1+\exp-\left(\left(\theta_{ii(t)}T_{(t)}\right)-\left|2\eta_t\right|\right)} \qquad (6.17).$$

For a connection between two different dimensions, i to j, the following applies;

$$\alpha_{ij(t)} = \frac{1}{1+\exp-\left(\left(\theta_{ij(t)}T_{(t)}\right)+\left|2\eta_t\right|\right)} \qquad (6.18).$$

The variable η_t is either the uncertainty regarding the decision, in the pre-response phase, or the teacher signal or error in the post-response phase. The details of how this value is determined will be discussed below, but it is proportional either to the discrepancy between output and feedback, or the uncertainty regarding the decision to be made. Its

different use in the two equations supports a system whereby sampling becomes increasingly distributed across the stimulus dimensions as uncertainty or error increases.

For the recurrent connection, the size of the level of uncertainty or error reduces the effective connection strength, decreasing the probability of consecutive sampling of the same dimension. For the connection between different dimensions, the value increases the effective connection strength.

Modulation of the effective strength according to the number of steps elapsed since the trial began is effected by the value of T_t. This value is determined by the following,

$$T_t = \frac{1}{1+e^{-g_T(t-\rho)}} \qquad (6.19),$$

where g_T is a gain on the function determining T, and ρ is a bias such that $T_{t=\rho}= 0.5$. The values of these parameters were set at $g_T = 0.25$, and $\rho = 25$. This means that early in the trial, T is close to zero such that almost none of the learned transition weight is involved in the sampling process. As the trial progresses the values of the transition weights become increasingly important in sampling. By the time the decision is made the value of T is close to unity.

The use of equations 6.17 and 6.18 to determine the effective connection strengths was motivated by, mostly, practical concerns. It is important to note that while this model involves the representation of information accumulating for a decision process over time, it was not specifically designed to model performance measures such as reaction times or differences in performance subject to time constraints. Testing of the model with respect to this type of task is beyond the scope of this thesis and remains an area for future research.

The practical concerns which motivated the forms of equations 6.17 and 6.18 were two-fold. Firstly, one of the problems which appears to be experienced by the previous dimensional attention models is slow early learning, particularly for the type VI task. One of the reasons for this is that the early development of associative weights on lower dimensionality sources tends to attenuate the activation of higher dimensionality representations.

While the early associations with low dimensionality sources are, in the long run, spurious, it takes time for the model to learn that this is the case. If one particular dimension is highly correlated with the first few feedback presentations, the transition

weights which develop will tend to reflect this, increasing the rate of sampling for that particular dimension at the expense of other dimensions. As a consequence the activation of configural representations, dependent on these 'ignored' dimensions, is reduced.

The inclusion of the error or uncertainty 'bias' to the effective connection strength attempts to alleviate this problem by generally increasing the rate of configural activation when uncertainty or error is high. Uncertainty and error will generally be at its highest during early trials such that this parameter may enhance configural activation most at this time.

Another problem experienced by the ATM model concerned the learning of exceptions in the type V structure. Exceptions to a rule must be learnt using representations with a higher dimensionality than the rule itself. When an exception is present, uncertainty will generally be at a higher level than it would be for a central category member. It was hoped that the extra uncertainty present on these trials might enhance the rate at which higher dimensionality representations were activated and thus enhance, to some extent, the rate at which exceptions might be learnt.

The second motivation behind the forms of equations 6.17 and 6.18 was to allow, early in the trial, some opportunity for all sources in the stimulus to achieve some level of activation prior to the decision being made. Diminishing the influence of the θ weights on the effective connection strength, according to the value of T, was a practical means of doing so.

6.3.2. Feedforward functions and channel activations

Like the two previous models, this model bases the probability of a particular label being selected on the logistic transform of the sum of the channel outputs. The measure is used throughout the trial and so is indexed by which step, t, the model is at as follows,

$$p(A)_{(t)} = \frac{1}{1+e^{-g_t \sum_c o_{c(t)}}} \qquad (6.20).$$

The gain in this case is indexed g_t to differentiate it from other gains used in the model. As with the previous models p(B) = 1-p(A). The actual output decision, or rather the probability of the decision, is taken at t = 50.

This model uses channel 'activation' rather than the probability of the channel being active. The output from a channel, c, to the decision node, l, at time t, or $o_{cl(t)}$ is as follows,

$$o_{cl(t)} = a_{c(t)} \sum_{\substack{s \in c \\ t \in t}} a_{s(t)} w_{sl(t)} \qquad (6.21).$$

Where a_s refers to the presence or absence of the cue or cue configuration detected by the source node, as described for previous models. The activation of the channel, a_c, is based on the 'history', across the trial, of sampling of the channel's dimension or dimensions. The indexing of a_s and the weight w_{sl} with **t** reflects that neither of these values change during the trial. The activation of a channel at step t is calculated according to the following,

$$a_{c(t)} = a_{c(t-1)} + \left(\left(A_{c(t)} - a_{c(t-1)} \right) \lambda_a \right) \qquad (6.22).$$

Where λ_a is the rate of change in the activation, and A_c is evaluated as follows,

$$A_{c(t)} = \begin{cases} 1 \text{ if } (d_1, d_2 \ldots d_U) \in \{ s_t, s_{t-1} \ldots s_{t-(U-1)} \} \\ 0 \text{ otherwise} \end{cases} \qquad (6.23).$$

The elements of the first set in the top half of the function are the dimensions of c, (u_c). As with the ATM model, the channel is conceptualised as the set of its U dimensions. The second set is the sequence of sampled dimensions, s, from step t back to t-(U-1) or the last U dimensions sampled. Note that the activation of a channel is set to zero at the beginning of each trial.

6.3.3. Alteration of associative and transition weights

The θ weights in this model are more complicated than in the last model and change differently according to what stage of the trial they are in. Rules regarding their alteration also vary according to whether they are recurrent weights or weights for transitions between different dimensions.

These weights are composed of two components. One component, $\Theta_{ij(t)}$, remains the same throughout the trial and is updated at the end of the trial. The other component, $\phi_{ij(t)}$, alters during the post-response part of the trial subject to whether the transition it refers to results in increases or decreases in error. The ongoing value of θ_{ij} at time t is composed of its two components as follows;

$$\theta_{ij(t)} = \Theta_{ij(t)} + \phi_{ij(t)} \qquad (6.24),$$

where t is a step in **t**. The change to the ϕ values occurs dependent on where in the trial the model is,

$$\Delta\phi_{ij(t)} = \begin{cases} a_{ij(t)}(|\eta_{t-1}| - |\eta_t|)\lambda_\phi & \text{if } 75 \geq t > 50 \\ 0 \text{ otherwise} \end{cases} \qquad (6.25).$$

Where $a_{ij(t)}$ is the activation of the transition at time t, that is $a_{ij(t)} = 1$ if $s_t = j$ and $s_{t-1} = i$, and equals zero otherwise. The rate of change is given by λ_ϕ.

The uncertainty or error, η, is again determined differently depending on whether the model is in the pre-response phase (uncertainty) or post-response phase (error),

$$\eta_t = \begin{cases} \delta'_{t-1} - p(A)_t & \text{if } 1 \leq t < 50 \\ \delta_t - p(A)_t & \text{if } 50 < t \leq 75 \\ 0 \text{ otherwise} \end{cases} \qquad (6.26).$$

The delta values for the pre-response phase depend on the label decision which is most dominant at the time. So $\delta'_t = 1$ if $p(A)_t > 0.5$, and equals zero if $p(A)_t$ is less than 0.5, such that B is the dominant decision. If $p(A)_t = 0.5$, then δ'_t is assigned a value of 0.5 such that error is zero. For the post-response phase of the trial, the delta values take on values determined by feedback such that $\delta_t = 1$ if the stimulus is a member of category A and $\delta_t = 0$ if the stimulus is a member of category B. At the end of each step, for the beginning of the next, the ϕ weights are updated,

$$\phi_{ij(t+1)} = \phi_{ij(t)} + \Delta\phi_{ij(t)} \qquad (6.27).$$

A point to note is that the uncertainty parameter, η, used in the pre-response phase is dependent on how discrepant the current output, at step t, is with the decision which was dominant on the last step. This will tend to result in increases to uncertainty if a transition leads to a change in the dominant decision. Even if there is no absolute change in certainty from one step to the next, a change in the sign of the summed channel output will result in a higher value for the uncertainty parameter. Given the role of the uncertainty parameter in equations 6.17 and 6.18, changes in the dominant decision are likely to decrease the subsequent probability of consecutive sampling of the same dimension.

The sizes of the changes in uncertainty (or error) are also controlled, somewhat, by the rate at which channel activation may change. Because the increments to channel activation are comparatively small in equation 6.22 (a value of 0.075 for λ_a was used), the increment rates have to be quite high for this model (4 used).

At the end of each trial, the transition weights are updated by altering the Θ_{ij} component of the transition weight. This update uses back-propagated error at the end of the trial, as well as the set of updates made to the ϕ_{ij} component during the post-response part of the trial. The updates are different depending on whether the transition weight is between two different dimensions or between a dimension and itself. For recurrent connections the value of Θ_{ii} at trial **t**+1, is calculated by the following,

$$\Theta_{ii(\mathbf{t}+1)} = \Theta_{ii(\mathbf{t})} + \lambda_{\Theta} \sum_{\substack{t=50 \\ t \in \mathbf{t}}}^{75} \Delta\phi_{ii(t)} \qquad (6.28).$$

This is simply the sum of changes to the ϕ component of the weight in the post response phase of the trial **t** multiplied by a rate constant λ_{Θ}.

For the connection between dimensions the update is more complicated,

$$\Theta_{ij(\mathbf{t}+1)} = \Theta_{ij(\mathbf{t})} + \lambda_{\Theta} \left(\left(\sum_{\substack{t=50 \\ t \in \mathbf{t}}}^{75} \Delta\phi_{ij(t)} + \sum_{\substack{t=50 \\ t \in \mathbf{t}}}^{75} \Delta\phi_{ji(t)} \right) + \left(b_{j(\mathbf{t})} - b_{i(\mathbf{t})} \right) \right) \qquad (6.29).$$

In this case, the change is a function of the changes to the ϕ component of the weight in the post-response phase for the transition i to j, but also for the transition in the opposite direction. The reason for this will be discussed in more detail below. Also added to the change is the difference between signals back-propagated to the origin and destination for the transition, as used in the ATM model. This back-propagated signal is calculated based only on the output on the final, 75th step of the trial **t**. The signal for a dimension d is calculated by a similar method to that used to determine the alteration of sampling probabilities, using equation 6.6,

$$b_{d(\mathbf{t})} = \sum_{d \in c} \eta_{(t=75)} o_{cl(t=75)} \qquad (6.30).$$

The use of the sum of the increments to the transition in the opposite direction to that of the weight being updated was a response to some 'difficulties' experienced by the model when just the increments for the transition in question were used. These difficulties

were interesting because they involved the development of cyclical transition weights, as noted for the type IV structure in the ATM model.

With the RATM model, cyclical sampling tended to occur within about three to four blocks for the type VI structure. Owing to the somewhat noisy nature of the sampling process, it is somewhat inevitable that for some trials a particular direction of transition will dominate. When this happens, particularly for the type VI structure, each transition in the dominant direction is likely to decrease error, by enhancing the activation of the fully valid three-dimensional channel. This will further enhance the directionality of the process as the trial progresses.

At the end of the trial, this pattern will be substantially reinforced such that it begins with a higher probability on the next trial. This is not necessarily the case for structures III and V where, on some trials, some directional transitions are highly likely to considerably increase error. The result is that while the learning on the type VI task begins characteristically slowly, once the cyclical pattern has been learnt, its performance quickly overtakes that on tasks III and V and, occasionally, type IV as well.

This is because, for the type VI task, the activation of the three-dimensional channel with cyclical sampling tends to approach unity quite rapidly. The learning of its association weights, being dependent on channel activation, thus tends to be more rapid than learning on the association weights for the types III to V tasks, which rely on a distributed pattern of activation across a number of channels.

Introducing the increments to the transition in the opposite direction as in equation 6.29 solved this problem. Whether this solution is particularly satisfactory will be discussed in more detail later.

Association weights are updated using a similar rule as that given for the previous two models. Each weight is updated at the end of a particular trial based on its channel activation and the error occurring at the end of the trial,

$$w_{sl(t+1)} = w_{sl(t)} + \left(\eta_{t=75} a_{s(t)} a_{c(t=75)} \lambda_w\right) \quad \text{where } s \in c \qquad (6.31),$$

where λ_w is the learning rate for these weights.

6.3.4. Simulation results and discussion

As with all of the previous models the model was tested twenty times on each of the six Shepard *et al.* (1961) category structures. Each run used new randomised orders of input pattern presentation. The model parameters used for the results presented below are given in table 6.1. Note all weights, including transition weights were initialised with a value of zero.

parameter	value	function
g_T	0.25	rate at which time elapsed in trial affects transition matrix (eq. 6.19)
ρ	25	number of steps before 0.5 of transition matrix is used (eq. 6.19)
g_I	3	gain on decision function (eq. 6.20)
λ_a	0.075	rate at which channel activation changes (eq. 6.22)
λ_w	1	associative weight learning rate (eq. 6.31)
λ_ϕ	4	within-trial transition weight learning rate (eq. 6.25)
λ_Θ	2.5	between-trial transition weight learning rate (eq. 6.28 & 6.29)

Table 6.1: Parameter values used for the results presented

The average p(correct) per block on the six tasks is shown in figure 6.17 and displays a similar pattern to that shown by the previous two models. The model exhibits the experimentally observed relative order of difficulty with the type II structure learnt with fewer errors than the types III to V.

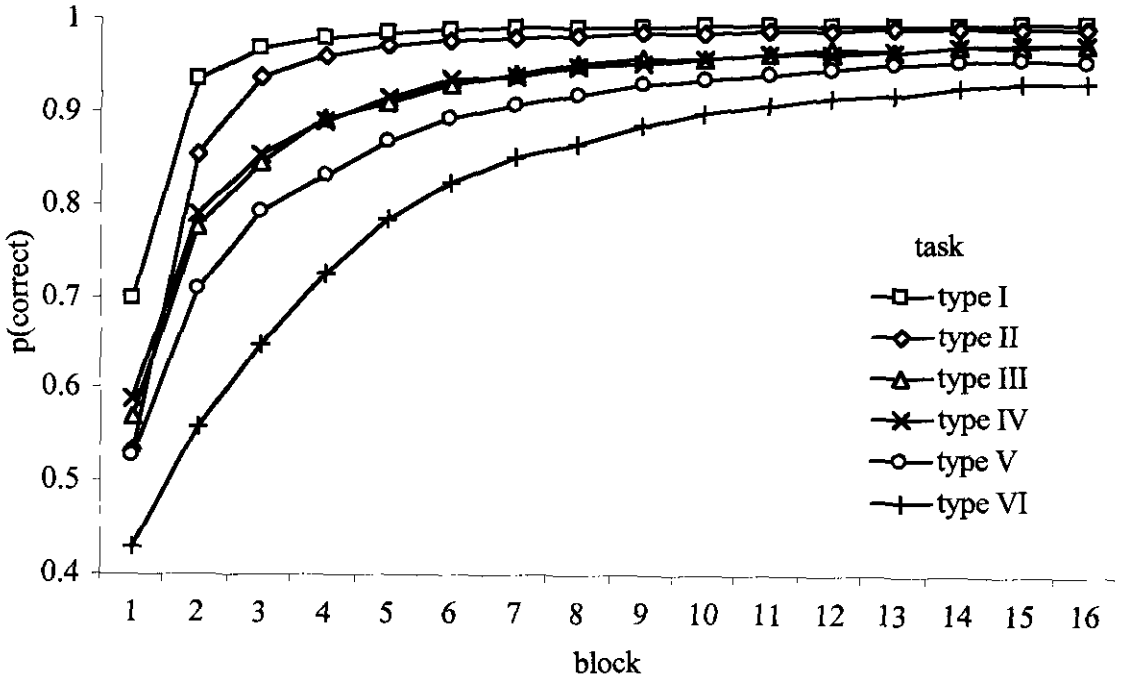

Figure 6.17: Mean p(correct) taken at t=50 of each trial per block, averaged across the twenty runs through each task for the RATM model.

As with the ASP model, and less clearly with the ATM model, there is a noticeable difference between performance on the type V structure and performance on the types III and IV structures. Like these models, figure 6.18 illustrates that much of this difference may be attributed to the low rate at which exception members of the type V structure are learnt.

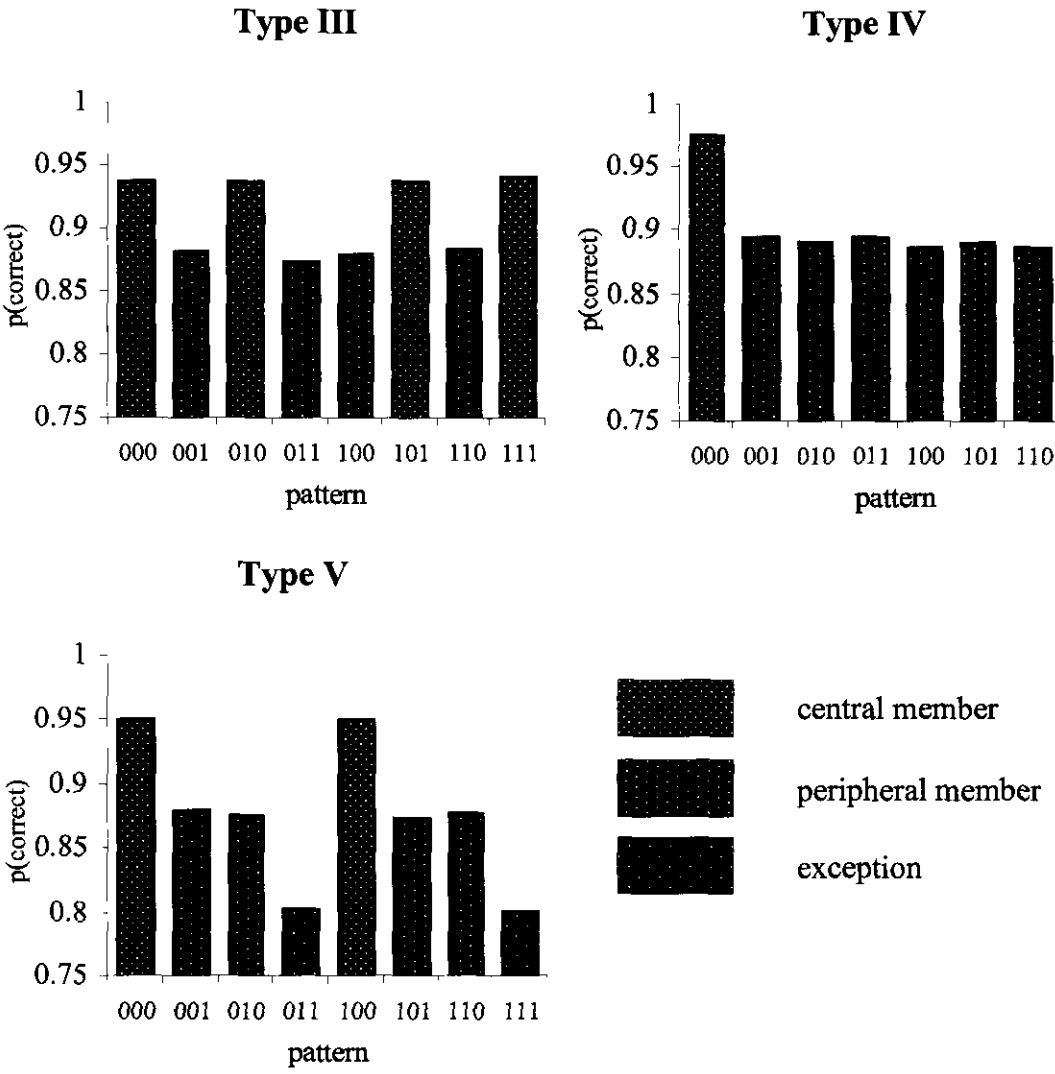

Figure 6.18: Performance of the model on individual patterns for category structures III, IV, and V. Performance is indexed in terms of the average probability of correct responding across the entire 16 blocks of the simulation. Figure 2.2 shows the relationships between these three types of category member in each category structure.

Dealing with exceptions is quite a general problem for models which use dimensional attention. As was described in chapter 3, attempts to model experiments in which learning of exceptions is examined tend to make use of modular network architectures which, in effect, change the way in which a stimulus is represented depending on whether it conforms to a dominant rule or an exception. Because dimensional attention

models generally only have one set of dimensional attention parameters relating to the entire task, they may lack the 'degrees of freedom' necessary to accurately represent exception learning.

The effects of this model's way of implementing dimensional attention, in terms of individual channel activations, is shown in figure 6.19. These are the averages at each block of the individual channel activations at the point at which the choice probability is determined, e.g. t=50.

The pattern displayed is fairly similar to that shown by the previous two models. The attention learning appears to have operated as expected for the types I and II tasks, rapidly attenuating the activation of non-valid and valid but redundant channels as in the previous two models. An exception is that the activation of the three-dimensional QRS channel in the type IV task tends towards a lower asymptotic value. The model produced no 'cyclical' patterns of transition weights, due to the nature of equation 6.29.

As with the previous models, the interactive nature of the associative weight learning rule resulted in a lower average rate of activation for the QS channel, than the QR and RS channels in the type III task. This is due to the fact that the valid sources in the QS channel only ever occur in the context of central category members and, consequently, are accompanied by lower average error signals.

The associative weights that developed for this model on the various tasks, while not shown here, were broadly similar to those given for the ASP model in section 6.1.4.3. Once the associative learning rule is made dependent on the error relative to combined channel contributions, the interactive patterns shown for the ASP model, and displayed for the ATM model, may be expected. This includes the development of weights for apparantly non-valid sources in the types III and V structures, as discussed in section 6.1.4.3.

Certain differences between this model and the ATM model can be seen in terms of the asymptotic transition weights developed during learning. These are shown in figure 6.20 in terms of the values of Θ_{ij} passed through a logistic and normalised across each origin dimension. These are the average resultant values across the last four blocks of training.

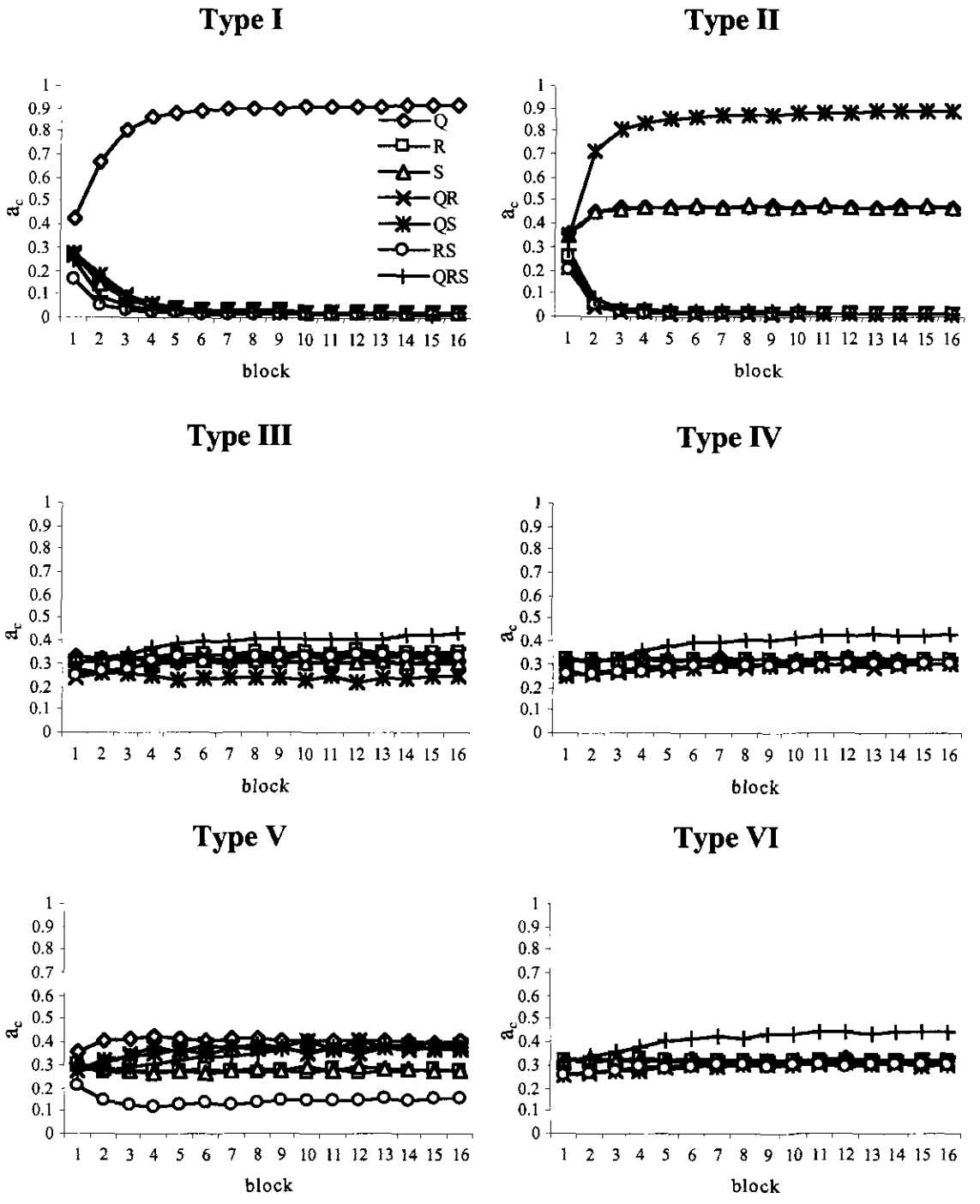

Figure 6.19: Average channel activations at point of decision, $a_{c(t=50)}$, per block. Key shown for the type I task applies to all graphs.

Comparing figure 6.20 with figure 6.14 reveals that the most noticeable difference is that recurrent connections on the irrelevant dimensions of the types I and II structures are not as low for this model as for the 'average' trial representation in figure 6.14.

This is primarily due to the absence of an interactive or competitive aspect to the learning rule for transition weights. The ATM model, using equation 6.16, took into account changes to the other transition weights from a particular dimension when evaluating the final increment. As learning progressed, less 'useful' transitions, regardless of their infrequency, would receive progressively larger, negative error signals.

In this model, transitions only tend to get incremented if they actually occur. The only competitive aspect of the learning function is the comparison of back-propagated signals used in the update of transitions between different dimensions, given in equation 6.29. Transitions between valid and non-valid dimensions are thus substantially reduced, despite the infrequency of transitions to non-valid dimensions. Recurrent transitions are wholly dependent on whether they occur or not. After a period of learning on the type I and II category structures the probability of recurrent transitions on non-valid dimensions will be very low as sampling tends to be restricted, mostly, to valid dimensions, particularly by the time the response is decided upon.

For the types III and V tasks, this lack of competition results in a more distributed pattern of sampling than that observed for the ATM model. Comparison of figures 6.13 and 6.19 illustrates that the present model maintains a higher level of activation in its less valid channels than does the previous model. This *should*, in theory, enhance the performance of the model on type V exception members by increasing the probability of three-dimensional channel activation.

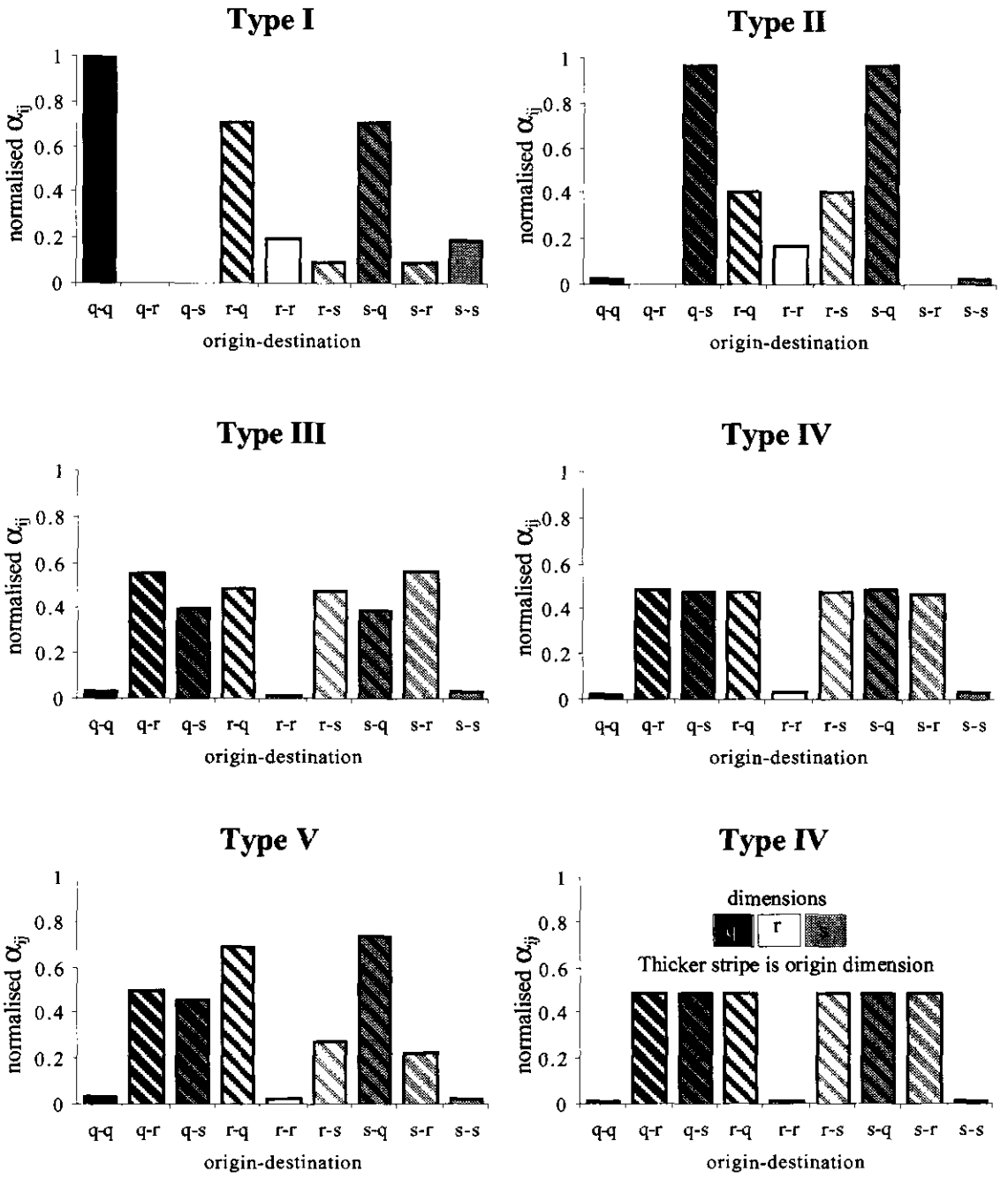

Figure 6.20: Average effective transition weights (α_{ij}), based on the values of Θ_{ij} averaged across the last four blocks of training, normalised across each origin dimension, e.g. q-q + q-r + q-s = 1. This is effectively the conditional probability of the destination being sampled at time t+1 given that the origin has been sampled at time t. Key shown for type VI applies to all charts.

Comparison of figures 6.12 and 6.18 indicates that while the central and peripheral members of the type V structure are learnt with almost the same ease for both models, there is a slight improvement for exception members in this model. The lack of competition in the current model also allows marginally superior performance on the central members of the type III structure, compared to that on the previous model shown in figure 6.12. Peripheral members, for both models, appear to be learnt at approximately the same rate. Comparing figures 6.13 and 6.19 illustrates that the QS channel, valid only for central category members on this task, is more active in the current model than in the last.

6.4. General discussion of sequential sampling dimensional attention models and the Shepard *et al.* (1961) tasks

The three models described in this chapter offer a means of implementing some form of 'dimensional selective attention' using a configural-cue form of representation. The first two models provide two distinct methods of implementing this feature which were both, to some extent, successful given the task provided. The third model represented an attempt to implement in a more specifically connectionist fashion the second model's adaptive transition matrix approach.

While there are clearly a number of conceptual issues regarding the models, this section will concentrate on how well the models seem to apply to the representation of learning of the six structures used by Shepard *et al.* (1961). Conceptual issues regarding the generalisability of the sequential sampling model of representation, the representation of the time course of a trial, dealing with more than two categories, and exception learning will be discussed in the next two chapters.

6.4.1. Persistent problems with the dimensional attention models

6.4.1.1 Poor early performance on the type II structure

One problem, which is suffered by all models in this chapter and all of the models in the last chapter, concerns performance on early blocks. All models appear to attenuate performance on the type II task for the first block to a lower level than that shown for types III and IV. For the experimental data of Nosofsky *et al.* (1994) (see figure 2.3), the fact that the type II task is easier than the types III and IV is manifested in the very first block of training.

Kruschke and Bradley (1995) have suggested that human performance on early learning trials is something simple networks may have a problem representing. Their suggestion is that additional processes such as short-term memory or strategic guessing may be at work early in learning.

For the models presented in this chapter attempts were made to enhance the rate of learning for both type II and type VI tasks, by enhancing the relative rate at which configural detectors were activated by the sampling process. As might have been predicted the success of these methods was mixed. While the rates of learning for types II and VI structures were enhanced, so too were the rates for types III, IV, and V, such that the (wrong) first block order was maintained. These structures are just as dependent on configural representations as the types II and VI.

As described above, for the configural-cue form of representation this problem emerges from the accumulation of weights on high frequency one-dimensional sources. These weights, for the type II structure, interfere with the attention learning aspects of the model causing attention to shift away from dimensions which are contributing error. This reduces the size of the error signal delivered to these 'erroneous' weights. Where dimensions ignored as a result of this process are part of a useful configuration, subsequent activation of that configuration is attenuated.

The version of ALCOVE implemented by Nosofsky *et al.* (1994) does not suffer from these problems. When unfamiliar stimuli are presented to the model it generates its response probabilities according to how similar the stimulus is to stimuli which have been presented before. The activation of these 'familiar' exemplar representations is likely to be much less than the activation of the 'new' exemplar. As such, while generalisation is likely to result in the model having a higher probability of being incorrect than correct (the nearest neighbours of exemplars in the type II are, in two thirds of cases, in the other category), the probability is likely to be close to 0.5.

Following exposure to all eight stimuli, after the eighth trial, none of the representations will have positive weights for the wrong category. Those activated by generalisation will generally contribute less to the response strength than the *actual* exemplar representation and, as such, the model will stop producing p(correct) values of less than 0.5 after this point. There is no chance that the model may predict the right

category by chance alone beyond the eighth trial. In addition, once all of the exemplars have weights 'pointing' in the correct direction, attention learning will be rapid. The attention learning in ALCOVE favours strongly the learning of tasks where one or more dimensions are wholly irrelevant to the task. As discussed in section 3.3.3.2.2 learning when one or more dimensions are irrelevant is accelerated by a process of recruitment.

In order to improve the fit of the ATM model to the first block experimental performance one may have to propose additional mechanisms. One of these might be some form of short-term memory. What might be desirable for the ATM model may be for it to evaluate, early in learning, whether *dimensions* are relevant or not. To do this the model would have to have access to the 'predictions' which might be made if the input were different to that currently experienced.

ALCOVE actually has this access in that the predictions made by all exemplars activated by generalisation are available to the learning process. As described in section 3.3.3.2.2, attention learning in ALCOVE is largely dependent on this generalisation-based activation. This form of access to 'absent' stimulus representations is not available to the configural-cue model. Here the only parts of absent stimuli that may be represented on a given presentation are those shared with the current stimulus.

Kruschke and Bradley actually implemented a model of short-term memory processes as an enhancement for a basic component cue network (Kruschke & Bradley, 1995). This type of network was unable to account for early training performance on a variety of tasks. Allowing it the ability to use information about previously presented, but currently absent, stimuli enabled improved fits to experimental data. Interestingly, ALCOVE does not seem to require any such mechanism to allow it to achieve a decent fit to early learning in the Shepard *et al.* (1961) tasks. As discussed above, this is probably because the model, owing to its mode of representation, already makes this information available to learning processes.

As discussed in chapter 3, the exemplar form of representation is not particularly useful in accounting for data involving stimuli that have different numbers of dimensions or components. The models that are capable of representing these kinds of relationships are models such as component and configural-cue models. If, as suggested by Kruschke and Bradley, short-term memory processes are required to allow these models to accurately

account for early training performance, it seems likely that the same requirement may obtain for the representation of learning in the Shepard *et al.* (1961) tasks.

More research would appear to be required to clarify whether the generalisation behaviour of exemplar representations provides an adequate source of the information to allow the modelling of observed early performance. It may be the case that other forms of representations used in a model, incorporating some other means of making use of absent stimulus information, may be just as appropriate for representing the fine details of early learning.

A constraint on the role of this process, however, appears to be suggested by both the human data and by the success of the ALCOVE model in representing it. It would appear from these sources that enhancement of early learning is not totally indiscriminate but, instead, may favour most those tasks for which some dimensions are irrelevant. The indication may be, therefore, that at least some of this facilitation of learning may be a function of the attention learning system.

6.4.1.2 Poor asymptotic performance and the handling of exceptions

All of the dimensional attention models presented in this chapter suffered somewhat from poor asymptotic performance, particularly for tasks III, IV and V. These tasks are characterised by the fact that the validity of dimensions and configurations of dimensions varies, depending on the stimulus presented. This poses certain problems for limited capacity dimensional attention models such as the three models in this chapter.

If, on a particular trial, feedback reveals that one dimension is more relevant than another, the attention learning process will tend to result in changes to the subsequent sampling probabilities for the dimensions, such that the less relevant dimension has a lower sampling probability than the more relevant one. For the models presented here, this means that activation of all sources dependent on the less relevant dimension will be attenuated relative to those dependent on just the more relevant dimension.

For the types III to V tasks, the dimensions that are relevant and irrelevant change according to the stimulus being presented. The result for these models is that performance on a stimulus is somewhat dependent on whether the dimensions that are relevant for it are the same as the dimensions for the stimuli which preceded it. If they are not, then one might expect high error signals and a change to the sampling process that favours the

dimensions relevant to the current stimulus. This tends to result in a persistently unstable set of sampling probabilities (for the ASP model) or transition matrix (for ATM and RATM models). This may attenuate average performance on all stimuli.

These problems are related to the inability of dimensional attention models, such as ALCOVE (Kruschke, 1992) to represent human learning performance in rule-plus-exception structures (Nosofsky, Palmeri, & McKinley, 1994; Palmeri & Nosofsky, 1995; Erickson & Kruschke, 1998), discussed in section 3.3.4.2. The learning of exceptions to low dimensionality rules requires attention to more dimensions than that used by the rule. The model must be able to identify exceptions if it is to be able to respond appropriately to them. It cannot do so if their defining dimensionality is ignored.

The tendency of dimensional attention models is to distribute attention across dimensions according to the relative frequency of their validity. This impedes learning of exceptions. If the defining dimensionality is not ignored, then learning of the rule is, relative to the learning of exceptions, attenuated to levels below that suggested by human data.

This difficulty will be returned to in the next chapters. Handling of exceptions may be particularly difficult for models which make use of rapid attention shifts, as some of the tasks to which they are suited seem to require an attention shift mechanism which would impede the learning of a rule-plus-exceptions structure.

6.4.1.3 Parameter settings in relation to early and late performance

The parameter settings adopted for use in these models were not formally optimised to increase the fit of the models to the data. One issue that seems relevant, however, is the high learning rates used for the attention learning processes. These high levels of learning rate were required to facilitate the early learning rates reported by Nosofsky *et al.* (1994), and illustrated in figure 2.3. It may well be the case that such high learning rate parameters contribute to the poor asymptotic performance of the models in relation to the types III to V structures.

It may be possible to improve the fit of all three models by incorporating some form of 'annealing' factor such as that used by Kruschke and Johansen in their RASHNL model (Kruschke & Johansen, 1999), mentioned in section 3.3.5.4. Such a factor is meant to slow down learning rates as the experimental simulation progresses. Applied to these

dimensional attention models, an annealing factor may facilitate improved asymptotic performance. Such a factor may be especially useful for the RATM model, which, like RASHNL, involves a rapid attention shift process. This kind of process may be particularly unstable under conditions where rules with low frequency exceptions are being learnt.

As was discussed in section 6.4.1.1, however, it may be the case that different processes may be at work during the early blocks of these experiments. While these processes may result in enhanced early learning, it may not be the case that they can be adequately simulated in terms of decreasing learning rates across the course of the experiment. Further research is clearly required to establish how learning progresses during the initial trials of experiments.

6.4.2 Sequential sampling models and dimensional attention

Despite these problems, the sequential sampling model does appear to offer a practical means of implementing some form of dimensional attention using a configural-cue approach to stimulus representation. The three models in this chapter indicate that this dimensional attention may be represented in at least two distinct ways.

As was discussed in section 6.1.4.4 there are certain conceptual and practical problems that might emerge from attempts to generalise the ASP model to other tasks. The ATM model, however, does not really represent a development of the ASP model. The ATM model is an alternative approach, which appears to begin without some of the problems of the ASP model.

It was decided to implement the ATM model rather than the ASP model in a more explicitly connectionist way, in the form of the RATM model, because the ATM approach appeared to be the most readily generalisable to other tasks. The tasks, which will be examined in the next chapter, involve learning about stimuli with different numbers of dimensions or components. In its current state, as discussed in section 6.1.4.4, the ASP model has no clear way of dealing with these problems.

Because the ATM model stores its attentional information *between* dimensions or components, its generalisation to this type of task is fairly straightforward. The initial sampling probabilities are always, simply, one over the number of dimensions. Where two dimensions or components have not been presented together, at the same time, one can assume an initial value for the transition weight of zero. If a dimension is absent from a

stimulus it may simply be assumed that there is no possibility of a transition to that dimension. Attention is learnt in terms of the relationships between dimensions and, as such, these relationships are only expressed to the extent that the dimensions which make up their components are present.

It is important to note that the rapid attention shift algorithm for the RATM model may also be applied to the ASP model. In this case one would simply 'send' a positive signal back to a dimension if the sampling of that dimension leads to a decrease in error and a negative signal if the sample increased error. It seems likely that one might also have to 'supplement' this process using back-propagated error signals in the same way as was required for both ATM models. The reason for this is that the error level change from sampling alone will have difficulty with regards to picking out the relationships involved in the type II structure.

For this structure, one dimension is irrelevant and the remaining two are relevant, but only in terms of the configural representations they support. A shift from the irrelevant dimension to one of the relevant dimensions is not likely to decrease error as, alone, the relevant dimension does not result in the configural activation required by the task.

For the ASP model, the result will be that positive signals will only be sent back to a relevant dimension when its sampling is immediately preceded by sampling of the other relevant dimension. During early training trials this is only likely to occur on, at best, half of the occasions on which a relevant dimension is sampled, thus attenuating early learning on this task. Whether the back-propagation part of the ATM's learning rule is important throughout learning or whether it simply represents another way of enhancing early learning perhaps requires further evaluation.

The actual use of the sequential sampling process to represent the activation of representations may be regarded as a somewhat speculative approach. It may be the case, for example, that it simply provides a way of simulating processes which actually operate in parallel. To some extent, this assertion may be supported by the fact that the directional learning of the utility of transitions had to be suppressed in the RATM model to allow it to simulate human performance.

Alternatively, it may be the case that information concerning the relative relevance of dimensions, for this type of experiment, is not stored in relation to the actual directions

of the transitions that revealed these relationships. Under some circumstances, however, directional sampling strategies may be stored. One example of this might be reading, where the components of words and sentences are likely to be sampled in a particular direction. It may be the case that where specific behavioural aspects to the sampling process are involved, such as eye movements, the directionality of error-reducing transitions may be remembered in some way. It seems likely, however, that the operation of selective attention processes is dependent on stimulus and context-dependent factors. Further research is clearly required to allow detailed theories regarding their role to be proposed and evaluated. Some of these issues will be returned to at the conclusion of this thesis.

The next chapter concentrates on applying a variant of the RATM model to other tasks. Of particular interest is the applicability of the model to tasks involving stimuli with different numbers of dimensions or components. This is a particularly relevant 'challenge' for the model, as these tasks generally seem to require completely different models of representation to those which have previously been successful for the Shepard *et al.* (1961) tasks.

Chapter 7. Further tests of the RATM model

The success of the ATM and RATM approaches with the Shepard *et al* (1961) category learning tasks suggested that dimensional attention could be usefully implemented using a configural-cue representation. This chapter will further explore the applicability of the model to other category learning tasks. The variant which this chapter will concentrate on is the RATM model. It was decided to apply this particular model to other tasks as it is the model that most specifically represents the processes of attention learning and representation involved in the adaptive transition matrix approach.

7.1. Base-rate effects

Employing as it does a form of rapid attention shift algorithm, it is suggested that the model might be usefully applied to tasks where this type of capacity seems to be important. Kruschke (1996a) suggested that rapid attention shifts might be essential for the representation of learning in certain tasks involving base-rate effects.

7.1.1. The inverse base-rate effect

As discussed in section 3.3.5.3, one experiment that seemed particularly amenable to this approach was Medin and Edelson's (1988) exploration of the inverse base-rate effect (*ibid.* experiment 1). Kruschke (1996a) simplified the experimental structure somewhat for his replication, reducing the number of categories from six to four and employing the abstract design shown in table 7.1.

As can be seen there are two rare diseases, two common diseases and a total of eight symptom sets or stimuli. For each disease, there is a symptom which perfectly predicts it. For each pair of common and rare diseases (e.g. C1 and R1) there is also a single, irrelevant symptom. The base-rate is such that common diseases occur three times more frequently than rare ones. As stated in Section 3.3.5.3, the inverse base-rate effect is observed on transfer trials given after a period of learning. When the combination of symptoms PC1+PR1 is presented, participants are more likely to assign the symptom set to disease R1 than to any other disease. Similarly presentation of PC2+PR2 is more likely to be followed by R2 assignments than any other.

Kruschke (1996a) suggested that the key role of base-rates was to cause common categories to be learnt before rare ones. This learning would involve the commitment of

components involved in the common category to the prediction of that category. Kruschke then suggested that rare category members tended to be learnt in terms of their *distinctive* features, i.e. those features which had not already been committed to the response strength for another category.

Kruschke's explanation for the effect in terms of the design shown in table 7.1 is, therefore, that the common categories will tend to be 'predicted' by compounds of the I + PC symptoms, with total asymptotic associative strength being divided between the two components. The rare categories are predicted only in terms of the PR symptoms. When a transfer stimulus, PC + PR, is presented, the PR symptom will have higher magnitude associative weights than the PC symptom, promoting selection of the rare disease.

symptom						
I1	PC1	PR1	I2	PC2	PR2	disease
1	1	0	0	0	0	C1
1	1	0	0	0	0	C1
1	1	0	0	0	0	C1
1	**0**	**1**	**0**	**0**	**0**	**R1**
0	0	0	1	1	0	C2
0	0	0	1	1	0	C2
0	0	0	1	1	0	C2
0	**0**	**0**	**1**	**0**	**1**	**R2**

Table 7.1: Abstract design of Kruschke's (1996a) experiment 1, investigating the inverse base-rate effect. C = common disease, R = rare disease, I = imperfect predictor, PC = perfect predictor for common disease, PR = perfect predictor for the rare disease. A value of 1 indicates the presence of the symptom and a zero indicates its absence.

7.1.1.1 Adjustments required for the RATM model

The RATM approach is likely to make similar predictions to ADIT about this experiment. Because common category members are, generally, presented before rare ones, the transition weights between PC and I components are likely to reflect not only the equal validity of these two components, but also the validity of the configuration which is activated by any sampling transition between the two components.

Subsequent presentations of common category members will, therefore, result in a pattern where error is reduced by alternate sampling of the two components. This will lead to increases in the transition weights between the components and decreases in the recurrent transition weights.

Presentation of a rare category member will, provided it follows presentations of common category members, result in an erroneous assignment of the stimulus to the common category. The subsequent attention shift phase will respond to the fact that error increases whenever sampling shifts from PR to I, or from I to I, and decreases when it either shifts from I to PR, or from PR to PR. The result will be negative transition weights between PR and I, and between I and itself, with positive weights developing between I and PR, and PR and itself.

Presentation of PC + PR transfer stimuli is likely to result in greater activation of the PR component because PC to PC transition weights are likely to be negative and the PR to PR transition weights will be positive. In addition, because the model employs an interactive learning rule similar to Rescorla and Wagner's (1972) rule, the PR symptom is likely to have larger associative weights than the PC symptom.

In order to model the inverse base-rate experiment described above, a few adjustments need to be made to the architecture described in the previous chapter. One of the advantages of using configural-cue representations is that one does not need to change the model of representation in order to represent learning in tasks involving stimuli with different numbers of components. Some aspects of the way in which the model deals with stimuli with different numbers of components need to be clarified here, but the model is fundamentally the same with respect to representation.

An important difference between this task and the Shepard *et al.* (1961) task, that has implications for certain parts of the model, is the fact that each category decision is

between four, rather than two, alternatives. This necessitates changes to the decision function and also to the way in which the 'certainty' of the network is evaluated. This, in turn, has consequences for the learning functions used which also need to be addressed.

7.1.1.1.1 Representation and sampling scheme

The representation and sampling scheme was fundamentally the same as that used by the RATM model for the Shepard *et al.* (1961) task. In this case, as table 7.1 shows, only 10 representation, or s, nodes are needed for this experiment. These are the six individual stimulus components and the four configural representations which are possible during training; I1+PC1, I1+PR1, I2+PC2, and I2+PR2. While the transfer stimuli presented at the end of training involve far more combinations of the stimulus components, none of these are present during training. These combinations would not become active during training and, consequently, their contribution to response strength during the transfer phase, when they actually would be presented, would be zero.

All possible transitions between the components were represented. While only some of these would acquire learnt weights during training, all of them occur during the transfer phase and so need to be represented. The sampling process described in the previous chapter is more or less the same in this model. In this case it is *components* which are being sampled rather than dimensions.

The important difference is that components which are absent cannot be sampled. This necessitates alterations to the functions used in calculating the sampling probabilities. The probability of a component being sampled is the same as the probability for a dimension being sampled, given in equation 6.7 of the last chapter. Now, however, the 'activation' of the component D, a_D is dependent on whether the component is present, as well as on the sum of contribution from other active components,

$$a_D = \pi_D \sum_d \tau_d \alpha_{dD} \qquad (7.1).$$

The presence of the component D is now indexed by π_D which takes on a value of one if the component is present on a trial and zero otherwise.

It is important to bear in mind that no changes occur for transition weights where one or both components are absent. All changes to transition weights, between components i and j are effectively gated by the product of π_i and π_j.

The predictions of the model, with respect to these base-rate effects, are generally similar to those made by ADIT (Kruschke, 1996a). One might expect the rapid attention shift mechanism to support configural representation of the common disease symptom sets. This is particularly the case if, as is most probable given the base-rates, the common symptom sets appear in training before the rare ones. In this case the adaptive transition matrix model would have nothing to choose from with regards to the two components of the common disease symptom set. A configural representation is likely to be activated which will, on subsequent trials, result in positive signals with respect to the transition between the two components.

The transition weight learning rule, developed for the RATM model, *tends* to lead to greater increments to transmission weights between equally valid components than to transition weights connecting the dimension or component with itself. This is because the increment to the transition weight between dimensions is a function of the sum of the increments to transitions between the components in both directions. This may result in greater increases to these transition weights than to the recurrent weights for each component. If this occurs then, particularly in a two dimension or component situation, the activation of the configural channel on subsequent trials is likely to exceed that of the individual components.

The result of this is that the configural representation will receive larger associative weight increments than either of the components. This, in turn, will lead to a situation where recurrent transitions will lead to a net decrease in certainty due to the resultant de-activation of the configural representation, further enhancing the difference between recurrent transitions and transitions between the two components.

Presentation of a rare symptom set will result in a problem for the precise implementation of the RATM model described in section 6.3. The problem emerges from the inclusion of the decision error or uncertainty in the determination of the ongoing transition matrix, incorporated in equations 6.17 and 6.18. The motivation behind this feedback process in the context of the Shepard *et al.* (1961) tasks was an apparent requirement to enhance the activation of configural representations, particularly during early training trials. While it was hoped that this would improve the early performance of the model on the type II and VI structure, and also address certain problems with the

learning of exceptions in the type V structure, the results were somewhat equivocal (see section 6.4.1.2).

What appears to be required, according to Kruschke's (1996a) hypothesis regarding the role of base-rates, is that initial presentation of the rare symptom set results in a shift of attention towards aspects of the stimulus which are *distinctive* or, as described above, otherwise uncommitted. The ADIT model, described above and in section 3.3.5.3.1, performs this shift because doing so reduces the level of error at the decision process.

While a similar process would apply for the rapid attention shifts of the RATM model, the high level of error that obtains on this initial presentation will result in error feedback to the transition matrix. Because this feedback promotes configural activation, it will seriously reduce the extent to which the irrelevant symptom is ignored. This would require very large changes, due to learning, to overcome.

Not overcoming the error contribution to the matrix will lead to enhanced activation of the configural representation PR + I and, consequently, a large increment to the associative weight for this configural detector. If the increment for this weight is larger than that for the PR representation, then subsequently the transition matrix is likely to develop in ways that further enhance the activation of this representation. The end result is likely to be predominantly configural representations of both categories, which will compromise the model's ability to display the inverse base-rate effect. This is because the associative weights for PC and PR representations will have developed according to equal activation levels (as half of a configural presentation) but with base-rates favouring the common cue.

These problems will disappear if the ongoing error level fed back into the transition matrix according to equations 6.17 and 6.18 is removed. The equation determining the value of the effective transition weight between i and j at time t, or $\alpha_{ij(t)}$ is, for this model,

$$\alpha_{ij(t)} = \frac{1}{1+\exp-\left(\theta_{ij(t)}T_{(t)}\right)}\pi_i\pi_j \qquad (7.2).$$

Note that only one equation is required by this model for the representation of connections between different components, and those between a component and itself. The implications of this adjustment will be discussed in more detail in the final chapter.

7.1.1.1.2 Increased number of categories

The other changes to the model necessitated by this experimental design concern the choice functions and the determination of error for the network. As discussed in section 3.1.2.2 it is straightforward to generalise the logistic choice function to choices involving more than two categories. The probability of a particular category, l, being selected at time t, $p(l)_t$ is given by the following function,

$$p(l)_t = \frac{e^{g_l \sum_c o_{cl(t)}}}{\sum_l e^{g_l \sum_c o_{cl(t)}}} \qquad (7.3).$$

The gain on the choice function is given by g_l, and o_{cl} is the output from channel c to label node l, as given in the previous chapter by equation 6.21.

The impact of this change is perhaps greater for this model than in many other connectionist models. This is primarily due to the nature of the learning algorithms proposed and used for the RATM model. As discussed in section 6.1.3, the difference between response *probability* and feedback is used as the basis for most of the teacher and error signals for the models described in chapter 6. This includes associative weight updates, back-propagated signals, and the determination of error changes for use in the rapid attention shift process.

Having four categories to choose from means that the 'chance' probability of a particular decision, given no other associative strength, is now 0.25 rather than 0.5. This means that there is an asymmetry with respect to the size of teacher signals, in relation to positive and negative instances of a category. From an initial naïve state, the maximum negative teacher signal is now –0.25, whereas the maximum positive signal is 0.75.

One of the biggest problems for this model concerns the way in which uncertainty or error is represented with respect to the rapid attention shift process. For the two-category task in the previous chapters, a single error parameter may be used to describe the discrepancy of the decision probability from the feedback. The level of error would be equal with respect to both categories. For the four-category situation the level of error will be different for different categories. One is therefore faced with a choice as to how, and even whether, one should represent error using a single parameter.

It may well be preferable to represent error in terms of separate parameters for each alternative in the decision process. As will be discussed below, such a treatment may provide extra degrees of freedom for the model and facilitate its ability to represent certain experimental data.

Here, however, it was decided to attempt a simpler approach making use of a single error parameter for the whole decision process. In this way, it was hoped to minimise the extent to which the model is changed for the purposes of this experiment.

One candidate which might be considered for use as a single error term is the discrepancy between the presence of the category which has been presented and the choice probability for that category. This would be, basically, one minus the probability that the correct category was chosen.

This measure was ruled out due to the problem of asymmetry between error for positive and negative instances of a category. For the shift process to result in a dominance of the distinctive component of the rare symptom set, error must be significantly reduced, on initial presentation of the set, by enhancing the probability of transitions away from the irrelevant symptom. Even by reducing the activation of the irrelevant symptom's representation to zero, the lowest error the system can achieve on the first presentation of the set is 0.75. The initial changes to the transition weights are thus dependent on variation of this form of error between 1 and 0.75.

The operation of the rapid attention shift process for exhibiting the inverse base-rate effect, seems to require that as much 'priority' be given to shifting attention away from representations which predict the wrong category label, as is given to shifting it towards representations which predict the correct label. This is particularly important in the case of the initial presentation of the rare symptom set, where there is no positive associative strength with respect to the rare disease.

In order to address this issue, a somewhat speculative model was developed which enabled error to be expressed in terms of a symmetrical measure. In this case error is determined by only taking into account the choice probability for the correct category label and the highest *other* choice probability. To represent this the set, L, of possible decision alternatives, l, is partitioned into three subsets called G, H, and I. G contains the 'correct' response, *i.e. the response for which* $\delta_{l(t)} = 1$. *H contains those alternatives with the single*

highest value of $(1-\delta_{l(t)})p(l)_t$, i.e. the highest choice probability for an incorrect alternative at time t. As with the previous model, t refers to a step of the trial and **t** refers to the trial itself. H may contain more than one member. I is the complement of the union of G and H, i.e. I = (G∪H)' or L - (G∪H). This set may be empty.

In order to calculate the single error parameter η_t, for error at time t, the following was used,

$$\eta_t = 1 - \frac{v(G)_t}{v(G)_t + v(H)_t} \qquad (7.4).$$

The value $v(G)_t$ is given by $p(l)_t$ where $l \in G$. The other value in the equation is v(H). This value is given by $p(l)_t$ where $l \in H$. This is not given as a probability, because H may contain more than one member (with equal choice probabilities) and, as such, the probability of set H would be given by the sum of the probabilities of its elements. Values for $p(l)_t$ are provided using the softmax function given in equation 7.3.

The error value thus implies that the system is only dealing with one pair of alternatives at a time, the correct alternative and those incorrect alternatives with the highest choice probability. Because it only deals with the probability of a single element of H, rather than the sum of these probabilities (where membership is greater than one), the value of η_t is equal to 0.5 when v(G)=v(H). The error level for a situation where the associative strength delivered to the decision process is zero is, consequently, 0.5 rather than 0.75.

7.1.1.1.3 Weight alterations

It was decided to use this error measure throughout the learning algorithms for the model used here. This means that associative weights are only updated if they are connected to a member of set G or set H. Associative weights are updated at the end of each training trial according to the following,

$$w_{sl(t+1)} = w_{sl(t)} + \left(\delta_{l(t)} - v(X)_{t=100}\right) a_{s(t)} a_{c(t=100)} \lambda_w \qquad (7.5).$$
$$\text{where } s \in c,\ p(l)_{t=100} \in X,\ X = G \wedge H$$

Note that another adjustment made for the implementation of this model is that the trial is 100 steps long, rather than the 75 used in the Shepard *et al.* (1961) simulations reported in the last chapter. Because the membership of H may change across the trial, the only

weight(s) which are updated are those providing associative strength to members of G and H.

This, in turn, affects the way in which error is back-propagated to the component nodes. Error is only back-propagated at the end of the trial via choice alternatives which are members of G and H at the end of the trial. The back-propagated signal to a component node d at the end of trial **t**, or $b_{d(t)}$ is evaluated by the following,

$$b_{d(t)} = \sum_{\substack{l \\ p(l)_{t=100} \in X \\ X=G \wedge H}} \sum_{d \in c} \left(\delta_{l(t)} - v(X)_{t=100}\right) o_{cl(t=100)} \qquad (7.6).$$

Except for the qualification that transition weights involving absent dimensions are not changed at all on a trial, these weights are altered by the same rules as those used in the RATM model of the last chapter. In this case, the error parameter, whose increase and decrease governs within-trial updates, is given by equation 7.4. The point at which the post-response phase of learning begins is still t=50, with the decision being made on the 50th step. For this experiment the post-response phase lasts until t=100 rather than t=75.

7.1.1.2 The experimental simulation

Kruschke's (1996a) experimental design involved participants being presented with 12 blocks of training trials. Each block consisted of the eight symptom sets, shown in table 7.1, presented one at a time, in a random order. Participants were given around 30 seconds to make their decision, and then were presented with the symptom set plus the correct category label (disease) for a further 30 seconds.

Following training, participants were then presented with two blocks of transfer trials. Each block consisted of 18 patterns, presented in random order, consisting of individual symptoms and novel combinations of those symptoms. During this phase no feedback was presented.

For the data reported here, the simulations were carried out by generating 30 sequences of 120 training trials. Two transfer blocks were presented to each 'simulated participant' but the order was not randomised for the simulation. Because no learning takes place during transfer trials (learning parameters were all set to zero), the order of presentation of these trials is unimportant for the model. While Kruschke (*ibid.* p. 16) did present two training blocks to ADIT in his experimental simulations the model, with its

learning parameters set to zero for these trials, will produce the same choice probabilities on both blocks.

The situation is slightly different for the RATM model presented here. On transfer trials where only one symptom is presented, because the probability of sampling this component is unity, the model will produce identical choice probabilities each time these transfer inputs are presented. When the transfer stimulus has more than one component, however, the nature of the sampling process means that the model is likely to produce different choice probabilities each time the same multi-component stimulus is presented. For the transfer results presented below, the choice probabilities are thus determined by averaging across each of the values for each presentation of the stimulus. The actual transfer stimuli tested were; the six individual symptoms presented alone, the four PC + PR combinations, I1 + PC2, I2 + PC1, I1 + PR2, I2 +PR1, I1 + PC1 + PR1, I2 + PC2 + PR2, I1 + PC1 + PR2, and I2 + PC2 + PR1.

7.1.1.3 Results and discussion

For the results presented below, the parameter values used are shown in table 7.2. While there was no concerted effort to optimise these parameters they are different to those used in the Shepard *et al.* (1961) simulations reported in table 6.1 of the last chapter.

As can be seen, four parameters were altered for this model relative to that presented in the last chapter. The learning rate parameters for the transition weights have both been increased, the within trial shift being increased considerably. As discussed above, Kruschke's (1996a) theory regarding the role of base-rates in learning relies heavily on the rapid shift of attention to distinctive inputs for the rare symptom set, on its first presentation.

parameter	value	function
g_T	0.25	rate at which time elapsed in trial affects transition matrix (eq. 6.19)
ρ	25	number of steps before 0.5 of transition matrix is used (eq. 6.19)
g_I	2	gain on decision function (eq. 7.3)
λ_a	0.075	rate at which channel activation changes (eq. 6.22)
λ_w	0.5	associative weight learning rate (eq. 7.5)
λ_ϕ	8	within-trial transition weight learning rate (eq. 6.25)
λ_Θ	3	between-trial transition weight learning rate (eq. 6.28 & 6.29)

Table 7.2: Parameter values chosen for the simulation of Kruschke's (1996a) experiment 1 on the inverse base-rate effect using the RATM model.

For ADIT (*ibid.*), and this model, this shift must be quite considerable to prevent there being any substantial change in the associative weights from the irrelevant symptom. With ADIT, the irrelevant symptom is likely to have developed a positive connection towards the common disease. In the case of ADIT, associative weights develop according to the discrepancy between the raw associative strength and the label feedback. This feedback is zero if the category is absent, and one if it is present. In most cases, therefore the irrelevant symptom will have a zero weight for the rare disease and a non-zero positive weight for the common disease.

If the attention shift does not significantly reduce the effective activation of the irrelevant symptom representation, it is likely to develop a positive weight with respect to the rare disease, and the value of its weight for the common disease is likely to decrease. When the common disease is subsequently presented, dependent on the size of the change in weights for the irrelevant symptom, the rapid attention shift may move attention away from the irrelevant symptom and towards the PC symptom. This is likely to reduce the extent to which the common disease is represented by the configuration of symptoms and attenuate any inverse base-rate effect which might be displayed by the model.

The same is true for this model. It was therefore felt that increasing the transition weight learning rates might better represent the proposed comprehensive nature of these rapid shifts. The associative weight learning rate was decreased for similar reasons, in that

smaller weight changes might facilitate the development of 'clearer' trends in the transition weights. As it turned out, the results suggest that either the associative weight learning rate, or the decision gain (which was also reduced), could perhaps have been a bit higher.

Figure 7.1 illustrates the learning curves for the model across the fifteen blocks of training data. The two common categories are collapsed into a single graph, as are the two rare categories. As can be seen the model takes longer to learn the rare assignments than the common ones. This is in line with Kruschke's (1996a) experimental results. Across the last 5 blocks of training participants averaged 0.954 p(correct) for the rare diseases and 0.980 for the common diseases.

The figures for the model are 0.834 p(correct) for the rare, and 0.951 p(correct) for the common categories. Kruschke reports that ADIT, with parameters selected for best fit to the *transfer* data, produced 87% correct for the rare categories and 96% for the common categories. The fit of the RATM model to early training trials shows a pattern which is considerably worse than human performance. Kruschke's experimental data shows that first block performance for humans was 49% correct for the rare categories and 68% for the common categories. The RATM model fails to reproduce this early capacity, with actual first block figures being 23% for the rare and 37% for the common. ADIT suffered similarly in Kruschke's simulations, with 16% and 54% for rare and common categories respectively.

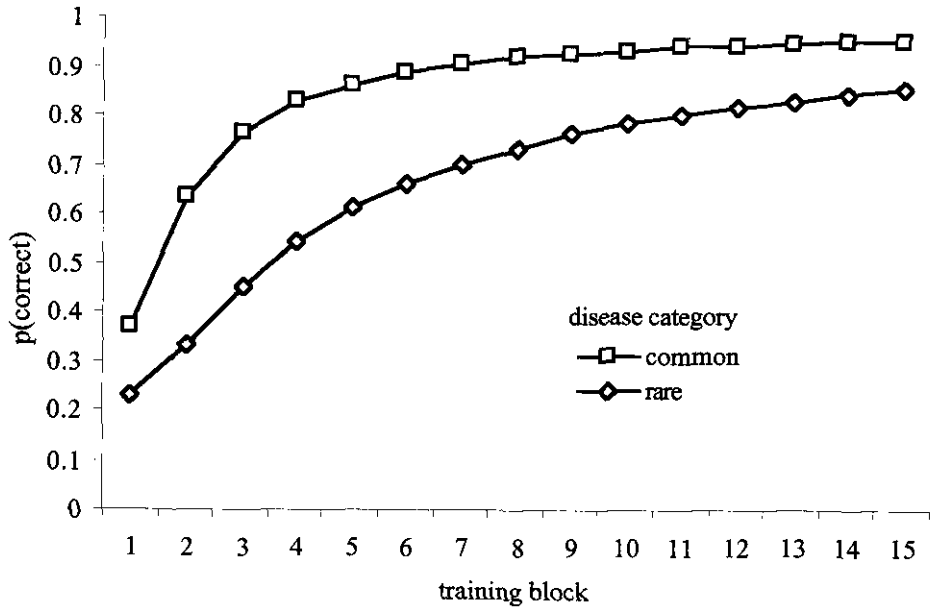

Figure 7.1: Average p(correct) of the RATM model for the simulation of Kruschke's (1996a) experiment 1 across the fifteen blocks of training data. Common and rare disease categories are collapsed into two graphs.

The early part of the learning curve for the RATM model shows much worse performance than that observed in humans. Across the first five blocks of training Kruschke's participants averaged 0.858 and 0.750 p(correct) for the common and rare categories respectively. The model produced proportions of 0.696 and 0.437.

As discussed briefly in section 6.4.4.1, it has been suggested that human performance on early learning trials is not generally captured by models using the delta rule or those based on the Rescorla-Wagner rule. Kruschke and Bradley (1995) have suggested that short-term memory processes or strategic guessing mechanisms, not generally represented by connectionist models of category learning, may play a role during early trials.

Without such processes it seems difficult to imagine how models such as ADIT or the RATM model presented here could possibly represent the early performance levels of human participants. Given that participants only see the rare category members once in the

first block, there is no way in which evidence capable of raising response probabilities above chance (0.25) can exist in this simple model, in the first block.

As discussed above, the main focus of this simulation is to investigate whether the RATM model might show the inverse base-rate effect. The human data reported by Kruschke (*ibid.* p. 6) is shown in table 7.3. It shows the mean choice proportions for the various sets of transfer stimuli. These figures were produced by averaging performance on each transfer trial across all participants. These averages are then collapsed across the different types of trial given in the table. So, for example, the choice proportions for C given for the symptom I in table 7.3 refer to average of p(C1| I1) and p(C2| I2) and those for Co for symptom I refer to the average of p(C2| I1) and p(C1| I2). There were two values for each transfer stimuli, for each participant, and so averaging was across these two trials for each simulation.

	choice proportions			
Symptoms	C	R	Co	Ro
I	0.746	0.174	0.049	0.031
PC	0.933	0.031	0.031	0.004
PR	0.040	0.911	0.018	0.031
PC+PR	0.353	0.612	0.022	0.013
I+PC+PR	0.580	0.402	0.013	0.004
I+PCo	0.406	0.080	0.469	0.045
I+PRo	0.219	0.085	0.031	0.665
PC+PRo	0.353	0.027	0.058	0.563
I+PC+PRo	0.719	0.036	0.036	0.210

Table 7.3: Experimental mean choice proportions for the transfer stimuli for the 56 participants in Kruschke's, 1996a, experiment 1 (p. 6). I = imperfect predictor for the two diseases, PC = perfect predictor for the common disease, PR = perfect predictor for the rare disease; C = common disease, R = rare disease, Co = the other common disease, and Ro = the other rare disease.

Kruschke noted a strong inverse base-rate effect for the PC + PR and PC + PRo transfer stimuli which, it can be seen, is shown by the simulation results in table 7.4. In

both cases, the disease associated with the rare symptom has a higher mean choice probability than the disease associated with the common symptom.

A small base-rate consistency was also noted for the I + PC + PR transfer stimuli (*ibid.*). This effect was amplified for the I + PC + PRo stimuli. Again, a similar pattern of choice probabilities is shown by the RATM model.

Comparison of the two tables illustrates that there is a fairly good qualitative fit of the simulation data to the experimental data with all of the first and second ranked choice proportions being accurately represented by the model. While some of the differences are somewhat attenuated in the model relative to the human data, it does indicate that the model would appear to be capable of reliably demonstrating the effects of interest.

Some of the differences between choice proportions may be increased by increasing the decision gain or the associative weight learning rate. Other differences are inevitable consequences of the architecture. The fact that the model's choice probability for the common diseases when presented with PC is less than p(R) when presented with PR, is a case in point. This is also characteristic of ADIT (Kruschke, 1996a, p. 17) and occurs because the common disease is predicted using configural information such that associative strength is distributed across several weights. This is not the case for the rare disease where the PR cue is responsible for almost all of the associative strength of the model with respect to the rare disease.

Factors which may affect the ability of this model to achieve closer fits to the data will be discussed in more detail below. Some of these are further highlighted by the performance of the model on the next task.

	choice proportions			
Symptoms	C	R	Co	Ro
I	0.555	0.240	0.103	0.103
PC	0.735	0.077	0.094	0.094
PR	0.057	0.789	0.077	0.077
PC+PR	0.254	0.503	0.122	0.122
I+PC+PR	0.512	0.340	0.074	0.074
I+PCo	0.301	0.196	0.384	0.118
I+PRo	0.220	0.157	0.096	0.526
PC+PRo	0.286	0.107	0.100	0.507
I+PC+PRo	0.708	0.069	0.059	0.163

Table 7.4: Mean choice proportions for the RATM model on the transfer stimuli for Kruschke, 1996a, *experiment* 1.

7.1.2 Base-rate neglect

As discussed in section 3.3.5.3, Kruschke (1996a) suggested that the same principles could be used to offer explanations of the inverse base-rate effect and the base-rate neglect noted by Gluck and Bower (1988a). He suggested that in base-rate neglect the normatively irrelevant feature is infrequent (relative to ther features) for the common category. As such, it is not as likely to be 'committed' to its prediction as the more commonly associated features. When the rare category occurs, this feature, being relatively uncommitted to the common category, may be associated with the rare category (Kruschke, 1996a).

The experiment that is simulated here is a variant of the previous experiment. In this variant Kruschke was trying to show that the inverse base-rate effect and base-rate neglect would be displayed within a single experimental structure (*ibid.* experiment 3). Table 7.5 shows the abstract structure of the experiment.

The experiment is divided into two substructures. The top half of table 7.5 shows the substructure designed to investigate base-rate neglect. As can be seen the third symptom set includes the symptom PRn. This means that PRn is not a perfect predictor for the rare disease, as now the conditional probabilities $p(Cn|\ PRn) = p(Rn\ |\ PRn) = 0.5$. As with Gluck and Bower's (1988a) paradigm, however, PRn is still the best available

predictor of the rare disease Rn. Base-rate neglect would, therefore, be demonstrated if participants chose Rn over other categories when tested with symptom PRn in isolation.

		symptom				
In	PCn	PRn	Ii	PCi	PRi	disease
1	1	0	0	0	0	Cn
1	1	0	0	0	0	Cn
1	1	1	0	0	0	Cn
1	**0**	**1**	**0**	**0**	**0**	**Rn**
0	0	0	1	1	0	Ci
0	0	0	1	1	0	Ci
0	0	0	1	1	0	Ci
0	**0**	**0**	**1**	**0**	**1**	**Ri**

Table 7.5: Abstract design of Kruschke's (1996a) experiment 3. C = common disease, R = rare disease, I = imperfect predictor, PC = perfect predictor for common disease, PR = predictor for the rare disease. Note n = base-rate neglect substructure, i = inverse base-rate substructure. A value of 1 indicates the presence of the symptom and a zero indicates its absence.

The bottom half of table 7.5 shows the inverse base-rate substructure. This substructure is the same as the top or bottom half of table 7.1.

7.1.2.1. Simulating the experiment

Due to the similarity of this experimental design with the last one, minimal changes were necessary for the RATM model. The only alteration which was necessary was to include additional configural nodes to represent the In + PCn + PRn and the PCn + PRn configurations contained in the training stimuli. The way in which Kruschke presented the stimuli to his participants was somewhat different for this experiment. In this case a single block of 21 transfer stimuli was presented to participants after every five blocks of training. The parameters used for this simulation were the same as those used in the last experiment shown in table 7.2. Again, 30 'simulated participants' were used.

7.1.2.2 Training results

Kruschke noted that his participants, as above, learnt the assignments for the common diseases significantly quicker than the rare ones. He also noted that they generally learnt the inverse base-rate substructure with fewer errors than the neglect substructure (Kruschke, 1996a, p. 9).

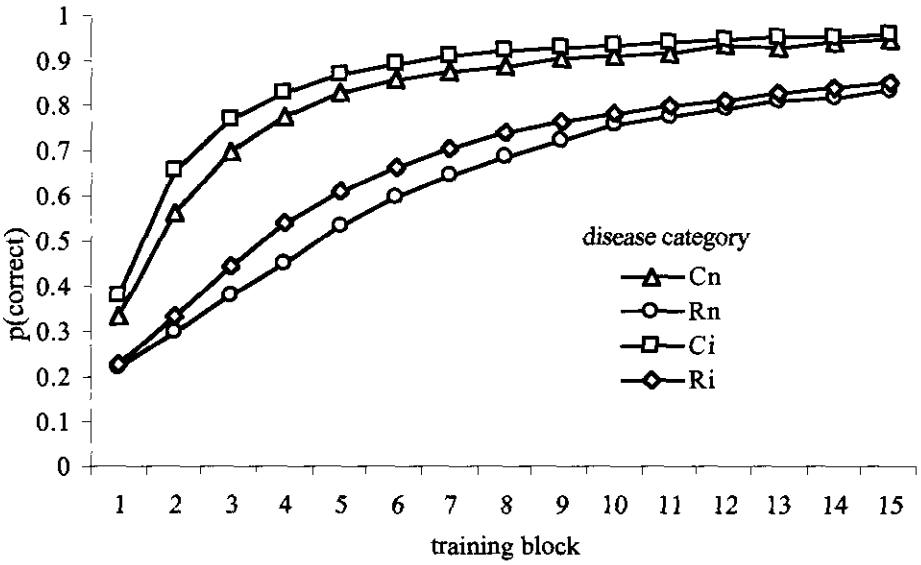

Figure 7.2: Mean probability of correct responding per training block for the RATM model on Kruschke's (1996a) experiment 3.

Figure 7.2 illustrates the learning curves for the RATM model, and qualitatively reflects experimental observations. Again, early performance is much worse than that reported for humans. The reasons for this are the same as those for the previous simulation. The model does display the observed differences in a qualitative fashion, with the neglect substructure being learnt with generally higher error probabilities than the inverse base-rate structure.

7.1.2.3 Performance on transfer stimuli

Table 7.6 gives the choice proportions of Kruschke's experimental participants on the transfer stimuli presented in the experiment. Transfer blocks were presented after every five blocks of training. Kruschke observed that there was no apparent difference in the onset times for the various base-rate effects, with both neglect and the inverse effect being evident in the first block of transfer data (*ibid.* p. 9). Consequently he decided to combine the data for the three transfer blocks. Table 7.6 thus shows the average across the three blocks for his 56 participants.

Base-rate neglect was noted in the results, with a strong preference for disease Rn when presented with the symptom PRn. The inverse base-rate effects observed in the previous experiment are repeated in this data. As can be seen there was a marked preference for the rare disease Ri when participants were presented with the conflicting pair of symptoms PCi + PRi.

As Kruschke points out, however, there is no inverse base-rate effect observed in the neglect substructure. When presented with the transfer stimuli PCn + PRn the result was a small, but not significant, preference for the disease Cn. It is also notable that participants preferred the Ri disease when presented with the PRn + PRi pair of symptoms. In addition Kruschke notes that the base-rate consistency effect, with respect to the Ii + PCi + PRi stimulus, is much attenuated (and non-significant) compared with its manifestation in the previous experiment. Kruschke tentatively attributes this 'dilution' to the effects of the three-symptom training pattern in the neglect substructure which, he suggests, may have predisposed participants to respond to any three symptom stimuli with the Cn disease (*ibid.* p. 10).

	Choice proportions			
Symptoms	Cn	Rn	Ci	Ri
In	0.637	0.268	0.030	0.065
Ii	0.036	0.060	0.780	0.125
In + Ii	0.359	0.174	0.365	0.102
PCn	0.833	0.113	0.018	0.036
PCi	0.036	0.030	0.893	0.042
PCn + PCi	0.488	0.054	0.411	0.048
PRn	0.131	0.774	0.042	0.054
PRi	0.006	0.030	0.030	0.935
PRn + PRi	0.024	0.292	0.036	0.649
PCn + PRn	0.542	0.400	0.018	0.042
PCi + PRi	0.012	0.024	0.321	0.643
In + PCn + PRn	0.875	0.107	0.012	0.006
Ii + PCi + PRi	0.089	0.024	0.482	0.405
In + PCi	0.327	0.155	0.500	0.018
Ii + PCn	0.536	0.048	0.345	0.071
In + PRi	0.232	0.089	0.030	0.649
Ii + PRn	0.084	0.506	0.265	0.145
PCn + PRi	0.293	0.030	0.006	0.671
PCi + PRn	0.077	0.458	0.423	0.042
In + PCn + PRi	0.696	0.036	0.000	0.268
Ii + PCi + PRn	0.137	0.173	0.643	0.048

Table 7.6: Mean choice proportions for participants on transfer trials from Kruschke's, 1996a, experiment 3.

Table 7.7 shows the performance of the RATM model averaged across the three transfer blocks for each stimulus. The major effects are demonstrated by the model in a qualitative fashion when compared with table 7.6. The model displays base-rate neglect with respect to symptom PRn and also shows the inverse base-rate effect for the symptom combination PCi + PRi. Somewhat unsurprisingly the model does not show the inverse base-rate effects for the neglect substructure. It shows a clear preference for the common disease, Cn, when presented with PCn + PRn.

Symptoms	Choice proportions			
	Cn	Rn	Ci	Ri
In	0.516	0.244	0.120	0.120
Ii	0.119	0.119	0.529	0.233
In + Ii	0.296	0.202	0.303	0.199
PCn	0.757	0.068	0.087	0.087
PCi	0.105	0.105	0.701	0.089
PCn + PCi	0.428	0.116	0.332	0.123
PRn	0.087	0.702	0.106	0.106
PRi	0.098	0.098	0.075	0.730
PRn + PRi	0.121	0.343	0.116	0.419
PCn + PRn	0.651	0.130	0.109	0.109
PCi + PRi	0.137	0.137	0.261	0.465
In + PCn + PRn	0.859	0.060	0.041	0.041
Ii + PCi + PRi	0.092	0.092	0.485	0.331
In + PCi	0.279	0.195	0.397	0.129
Ii + PCn	0.444	0.115	0.262	0.179
In + PRi	0.229	0.170	0.113	0.488
Ii + PRn	0.123	0.443	0.255	0.179
PCn + PRi	0.330	0.114	0.111	0.445
PCi + PRn	0.125	0.419	0.329	0.127
In + PCn + PRi	0.671	0.083	0.073	0.173
Ii + PCi + PRn	0.072	0.140	0.706	0.081

Table 7.7: Mean choice proportions for the RATM model on transfer trials from Kruschke's, 1996a, experiment 3.

This preference for the common disease on the PCn + PRn symptom set may be a little more clear than the preference shown in the human data in table 7.6. It might be suggested that the fact that this preference is very marked for the model is unsurprising in that the model has a unique configural representation for this transfer stimulus. During training this representation occurs in the context of the In + PCn + PRn symptom set, which is diagnostic of the Cn disease. As will be seen below, however, much of this preference may be attributable to the structure of the transition matrices which develop for these tasks.

The development of the transition weights for the inverse base-rate effect experiment described above (and for inverse-base rate substructure of this experiment) is a fairly straightforward process. The charts in figure 7.3 show the average, normalised, values of α_{ij} for the inverse base-rate substructure of the experiment. These values are determined by passing the values of Θ_{ij}, taken at the end of the last training block, through a logistic and then normalising them so that values for the same origin sum to unity. The values represented in figure 7.3 are the average normalised values across the thirty simulations.

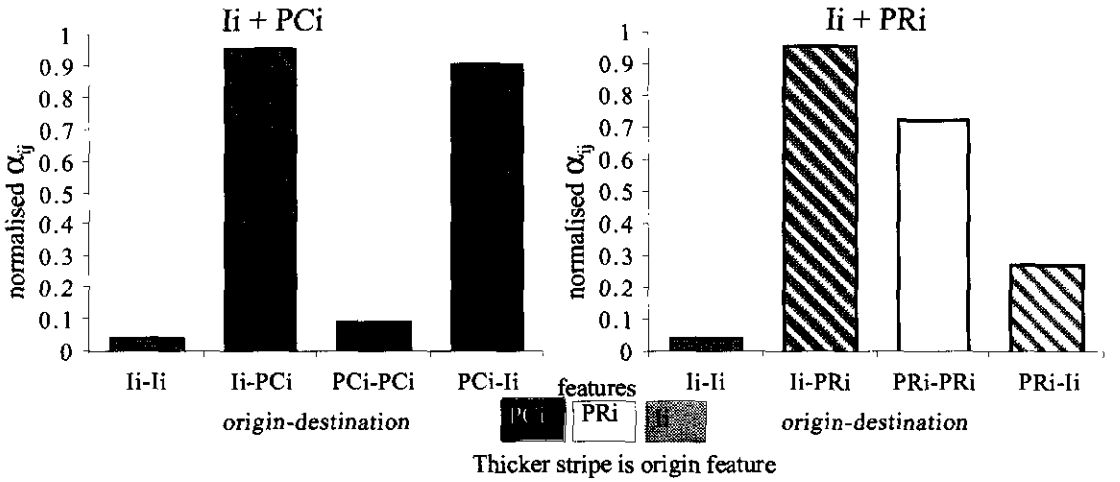

Figure 7.3: Average effective transition weights (α_{ij}) at the end of training, normalised across each origin component, e.g. Ii-Ii + Ii-PCi = 1. The two charts show the values for the two training stimuli Ii + PCi and Ii + PRi.

The difference between this experiment and the Shepard *et al.* (1961) experiment, for this model, is that not all of the transition weights are 'effective' for each stimulus presented. Figure 7.3 thus shows the effective transition matrix which obtains for the model on presentation of the common stimulus set, Ii + PCi, and the matrix which obtains on presentation of the rare set, Ii + PRi. These are the approximate average transition matrices as they do not take into account the value of T_t. This value will be close to unity at the point of decision, and so the above figure may represent the transition matrix quite well.

As can be seen the transition matrix for the common disease is quite different to that for the rare one. In the Ii + PCi chart the probability of consecutively sampling the

same component is quite low, whereas the probability of sampling the other component is high. This matrix will strongly support the activation of the configural representation (Ii, PCi) and result in the distribution of associative strength suggested by Kruschke's (1996a) hypothesis. For the rare stimulus, Ii + PRi, the matrix represented in the right hand panel indicates that while there will be some sampling of the Ii component, the majority of samples are likely to be of the PRi component. For this component there is a chance of recurrent sampling, which is generally absent for the irrelevant component.

For the neglect substructure the situation is different. The approximate transition matrices for the three stimuli presented during training are shown in figure 7.4. These values are calculated in the same way as those for figure 7.3.

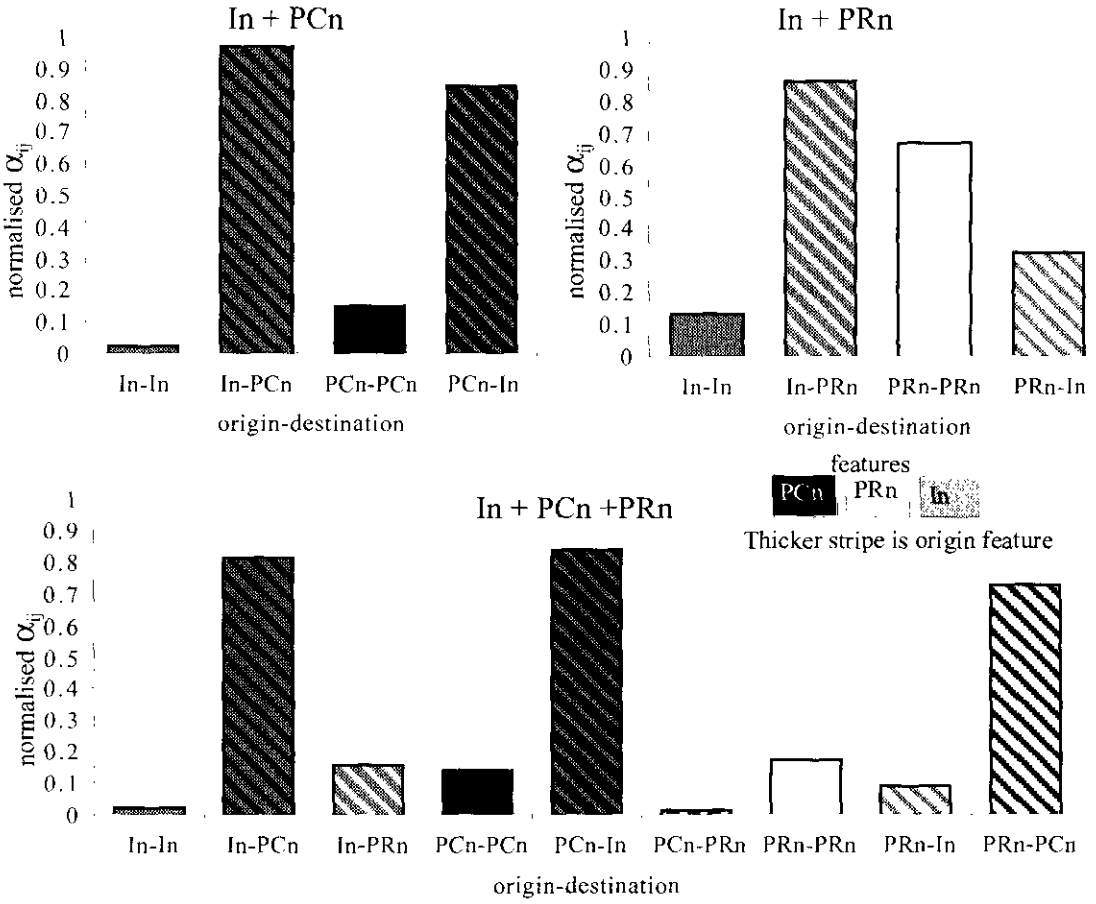

Figure 7.4: Average effective transition weights (α_{ij}) at the end of training, normalised across each origin component, e.g. In-In + In-PCn = 1. The three charts show the values for the three training stimuli In + PCn, In + PRn, and In + PCn + PRn.

The situation is somewhat more complicated for the neglect half of the substructure than for the inverse half. The top two charts in figure 7.4 display similar average transition matrices. For the In + PCn stimulus similar alternate sampling of the two components is promoted by the transition matrix. In this case there is a somewhat increased chance of consecutive sampling of the PCn cue compared to that shown in the left-hand panel of figure 7.3.

For the rare symptom set, again, the matrix appears to be quite similar. For this pattern, the PRn component is likely to be sampled more frequently than the In component. There is a somewhat increased chance, however, of consecutive In samples compared with that displayed in the inverse substructure.

The reason for the absence of the inverse base-rate effect from this half of the substructure is highlighted by the effective transition matrix which results upon presentation of the In + PCn + PRn symptom set, shown in the bottom panel of figure 7.4. For this stimulus there is evidence of a high transition weight from the PRn to the PCn components. The effect of this is illustrated by the effective, average transition matrices which occur in the presence of the transfer stimulus PCn + PRn shown in figure 7.5.

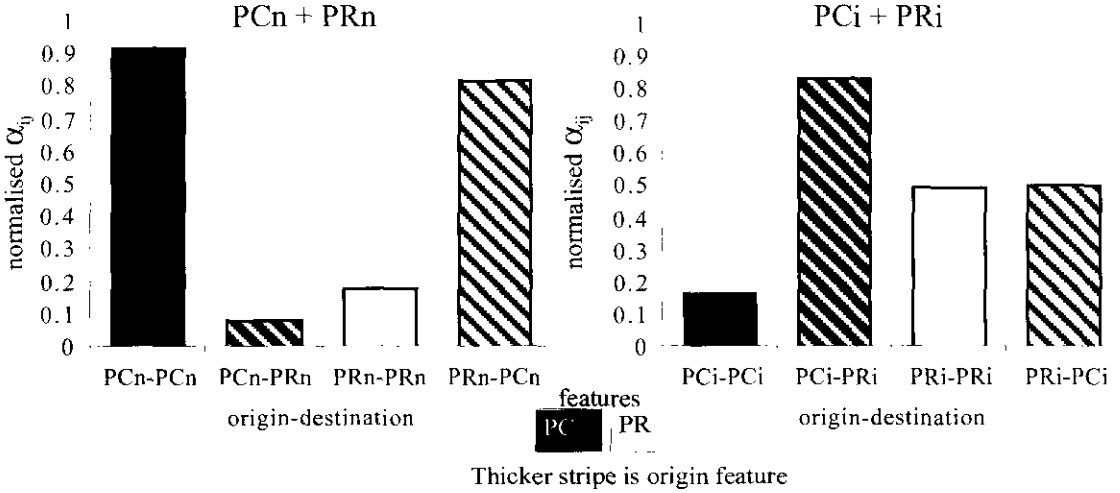

Figure 7.5: Average effective transition weights (α_{ij}) at the end of training, normalised across each origin component, e.g. PCn-PCn + PCn-PRn = 1. The two charts show the values for the transfer stimuli PCn + PRn and PCi + PRi.

These values were calculated in the same way as those for figures 7.3 and 7.4. In this case it is important to note that if the transition did not occur in training, i.e. the PCi to PRi and PRi to PCi transitions, then the value of α_{ij} for this transition was 0.5.

For the PCn + PRn stimulus, all of the transitions shown occur in training within the In + PCn + PRn symptom set. As can be seen, for this stimulus there are high probabilities of recurrent sampling of the PCn symptom. The PRn symptom shows a pattern whereby its sampling is most likely to be followed by sampling of the PCn component.

During presentation of the In + PCn + PRn symptom set, sampling of PRn, with its associative connection with the Rn disease, is likely to considerably increase error, this will result in transition weights *leading* to PRn being reduced. This relationship does not always hold for the In to PRn weight, as during presentation of the rare symptom set, sampling PRn is likely to decrease error. It is, however, *always* true for PCn symptom.

Similarly, during presentation of the In + PCn + PRn symptom set, transitions from PRn to PCn are likely to lead to reductions in the overall error, consequently the transition weight between these two components is likely to increase. The result is, as shown in figure 7.5, a matrix for PCn + PRn which is likely to promote high levels of activation for the PCn component and low levels of activation in the PRn component. Activation of the (PCn, PRn) configuration is likely to be fairly low. Thus the lack of an inverse base-rate effect for this substructure is not, principally, a function of the presence of a configural cue (although some contribution is likely to occur). The fact that the choice proportion for the Rn disease is quite a bit higher than that given in the experimental data, shown in table 7.6, may be a problem for the model as it would appear to be a fairly robust characteristic of the way in which learning and representation operates. Further investigation of the model and the experiment may be indicated to determine the importance of this characteristic.

In the inverse base-rate effect substructure, the matrix of transition weights is as might be expected. While there is no development of transition weights between the PCi and PRi components during training, the PCi to PCi weight is likely to be very low as indicated by the left-hand panel of figure 7.3. If this value is considerably less than zero, then when normalised along with the 'blank' transition weight for PCi to PRi, it will retain a low value and thus promote PCi to PRi transitions. The indication from this chart, is that

the PRi to PRi transition weight does not appear to acquire, on average, much in the way of a positive value. It would appear to be approximately equal to the 'blank' weight. This is enough to allow its activation to exceed that of the PCi component and, taking into account the larger size of its associative weights (as indicated by performance on the PCi and PRi transfer stimuli), thus facilitate the inverse base-rate effect.

Despite the general attenuation by the model of most of the 'preferences' given in table 7.6, there is a qualitative match by the model to the human data on all of the individual transfer stimuli. For all stimuli, the largest choice proportions are the same and, for most of the transfer stimuli, the second 'choices' of the model are also the same.

The model does appear to experience some difficulties with regards to the prediction of levels of differences between the choice proportions, both on this task and on the 'pure' inverse base-rate investigation described above. Some of these difficulties may be ameliorated by suitable parameter settings, but others may be more persistent functions of the architecture itself.

The choice probabilities for diseases in the other substructure when tested on single-symptom transfer stimuli are fairly high. In some cases these exceed the choice probabilities for the other disease in the same substructure. Although the differences between many of these 'improbable' choice proportions are unlikely to be significant for the human data, the pattern displayed by the model is indicative of a potential problem for the model.

One of the principal reasons for this pattern and, to a great extent, the attenuation of the size of differences in choice probabilities, is the reliance of the model on inhibitory associative weights. Symptoms are always negatively correlated with the presence of diseases from the other substructure. Ri never occurs in the presence of PCn, during training, for example. In this case there is likely to be some development of negative weights between, for example PCn and Ri and PCn and Ci. When In is subsequently presented in isolation as a transfer pattern, the 'loss' of these negative weights from PCn will enhance the probabilities of Ri and Ci selections.

In fact, being the most frequent symptoms, In and Ii in this model are likely to play a significant role in 'deciding' which 'half' of the structure is most likely. Its absence from

a transfer stimulus will therefore result in increased uncertainty about which half of the structure a disease is likely to belong in.

7.2 General discussion

Despite the shortcomings described above, the fact that the model can display the desired base-rate effects on transfer stimuli is somewhat encouraging for the RATM model and the sequential sampling model of representation. Significantly, there does not appear to be any single model capable of reproducing the order of learning difficulty for the Shepard *et al.* (1961) tasks *and* displaying base-rate neglect and the inverse base-rate effect. It may, of course, be argued that the differences between the model used on these tasks and the model described in the previous chapter mean that it isn't a single model at all. These differences, particularly the nature of the effective transition matrix under condition of error or uncertainty, are representative of an important problem in the modelling of categorisation. This issue will be returned to and discussed in more detail at the conclusion of this thesis.

One important component of the models that is the same is the sequential sampling model of representation. Models that have been successful with respect to base-rate effects have not used the exemplar model of representation, tending instead to employ component cue representations. These representations are, of course, not capable of modelling learning of the Shepard *et al.* (1961) tasks.

Whether the shortcomings of the RATM model described here could be ameliorated by a more concerted effort to optimise the parameters of the model awaits further research. It may be the case that more processes may have to be taken into account, in order to provide a comprehensive description of the learning and transfer effects observed. Investigation of the nature of these 'extra' processes is likely to require further experimentation to allow models with sufficient generality to be developed.

Despite the success of models such as ADIT (Kruschke, 1996a) and EXIT (Kruschke, in press a and b) in modelling the base-rate effects described above, their dependence on a component cue form of representation provides an obvious limitation to the approach. The experiments to which models such as ADIT and EXIT have been applied only involve the learning of category structures which do not require configural representations.

It seems unlikely that the effects that are well represented by ADIT and EXIT are confined to the types of experiments to which these models are actually restricted. Whether base rate-effects, such as those described above, may be displayed for category structures which actually need configural representations is something which requires further research. If the effects are observable, using stimuli which involve different numbers of components, then the RATM approach would appear to be the only model which can begin to represent them.

7.2.1 Kruschke's theory on base-rates and order effects

The RATM model may be described as an alternative method of formalising Kruschke's (1996a) theory regarding the role of base-rates in category learning (see section 3.3.5.3). This theory relies heavily on the order in which examples of the common and rare categories are presented. This reliance extends to both ADIT and the RATM model. The base-rates actually used in Kruschke's (1996a) experiments mean that, generally, members of the common categories are presented prior to members of the rare category. The inverse base-rate effect occurs, according to the theory, because people learn to predict the rare category in terms of the rare feature or symptom in isolation and the common category in terms of the common and irrelevant symptoms in conjunction.

The isolation of the rare component occurs because, even though the rare symptom is always presented alongside the irrelevant symptom, the irrelevant symptom has generally been committed to predicting the common disease. With ADIT, EXIT, and with the RATM model, the only way in which this effect can occur is if the common disease is presented prior to the rare one.

If the rare disease is presented first, then the irrelevant symptom will have equal associative strength to the rare symptom, with respect to the rare disease. Subsequent presentation of the common disease will result in 'attention' shifting away from the irrelevant symptom, such that more associative strength is allocated to the common symptom. One might expect, under these circumstances, to see some reversal of the inverse base-rate effect. The irrelevant symptom may, at transfer, result in predictions of the rare disease and the rare + common transfer stimulus may result in a common disease diagnosis.

This is certainly the case with the RATM model. 'Simulated participants' tend to display this reverse pattern when the rare disease is presented prior to the common one in training. Because the common disease is more likely to be presented before the rare disease, the average response probabilities, across simulated participants, display the inverse base-rate effect.

Kruschke (1996a) did not discuss whether the precise order of stimulus presentation in the first block, for his human participants, was related to subsequent manifestation of the base-rate effects at transfer. It would appear, however, to be something of interest, as both his theoretical account, and the ADIT model, predict that some relationship will exist. Kruschke has addressed this issue, to some extent, in a recent experiment which indicates that the inverse-base rate effect is preserved despite the introduction of a later training phase in which the base-rates of categories are reversed (Kruschke, in press b). Further experiments may be required, however, to determine whether base-rate effects are built into knowledge as early in training as suggested by attention shift models.

7.2.2 Time-scale of learning: how rapid *is* rapid?

ADIT and the RATM model are actually suggesting a lot more about why base-rate effects occur than is described by Kruschke's (1996a) theory about the role of base-rates in learning. These models provide a framework in which Kruschke's broad assertions about the role of base-rates may be seen as valid.

In the rapid attention shift models, learning is represented as a two-stage process in which associative weights between representations and outcomes are altered *following* a period of evaluation of the contribution of each representation to the outcome. Kruschke's theory cannot be currently modelled unless this two-stage model of the learning process is implemented.

The base-rate experiments described above, and the Medin and Edelson (1988) experiments on which these were based, all involved participants being allowed some fairly lengthy post-response exposure to both feedback and stimulus. As such, the two-stage model of learning, implied by rapid attention shift models, is not really being adequately tested by these experiments.

One might provide a more specific test for the models by varying the period of co-presentation of stimulus and feedback. The models would predict that removing the opportunity for this attention shift to occur might also remove the base-rate effects. Many experimental paradigms in associative learning research involve no co-presentation of stimulus and feedback: the stimulus is removed prior to feedback being presented (e.g. Pearce and Redhead, 1993). Single-stage associative learning models are often adequate to account for the learning effects noted in these experiments. Their success implies that some form of memory exists for the stimulus, allowing learning to be described 'as if' the stimulus was actually still 'activating' the representations between stimulus and response.

This may leave room for the two-stage models, such as the RATM model presented here, to suggest that the attention shift process may operate on similarly activated representations. As such, actually testing the two-stage theory may require more sophisticated procedures, such as the use of masking stimuli following stimulus offset.

The representation of learning as a two-stage process in models like the RATM and ADIT is actually fairly crude. There is no specific commitment, for example, to describing how long the attention shift phase lasts before associative weights are updated. It may be the case that this duration is dependent on some criterion of 'error minimisation' but, in this case, it seems difficult for the model to be able to identify exactly when error had been minimised. A criterion-based representation of the process may have to suggest that the onset of associative weight modification may be a stochastic process dependent on the rate at which error was being reduced by the attention shift process. In this scenario, as the rate of error reduction decreases, the probability that the associative weights will be updated increases. Whether this halts the attention learning process or not, would require further research and specification.

The assumption of the rapid attention shift models is that associative learning may be described as the last thing to happen before the next stimulus is presented. This, for a situation in which the inter-stimulus interval is restricted in some way, is suggestive of a model whereby the presentation of another stimulus halts the attention learning process and is followed by associative learning about the previous stimulus. The post-response time course of learning is something that is comparatively under-researched but, given the

development of theories and models which propose some structure for this phase, this may have to change.

7.2.3 Pre-response processes

The requirement for a *post-response* attention shift as the key to establishing the conditions for the manifestation of the inverse base-rate effect may not be essential. While some form of attention shift seems justified, *when* this shift occurs is not necessarily accounted for by the data. It may be possible that at least some of the shift takes place before the response is made.

The high average choice probability (see section 7.1.1.3) for the rare disease, upon initial presentation of the rare symptom set, suggests that some pre-response process is in operation, at least for these trials. Kruschke and Erickson (1995), and Kruschke and Bradley (1995), suggested that this probability, not predicted by ADIT, EXIT, or the RATM model, might be the result of some process of 'strategic guessing'.

Strategic guessing, also known as 'eliminative inference' (Juslin, Wennerholm, & Winman, in press), is, it may be argued, deployed by participants when faced with a novel stimulus. The response to an unfamiliar stimulus is to select an unfamiliar response (Kruschke, in press b). Whatever this process involves, it appears to enable participants to ignore the evidence in favour of the common disease being provided by the presence of the irrelevant symptom. As such, at least some of the enhancement of the rare disease's probability of selection may be accounted for by a shift of attention towards the novel symptom and away from the familiar, irrelevant one.

Why this shift should occur, and how it might be represented in a connectionist network such as the RATM model, is open to question. The fact that response probabilities for the unfamiliar diseases are elevated to a level far beyond the chance level given by totally ignoring the familiar stimulus, implies that attention shifting cannot be solely responsible for performance. It would appear that familiar categories are being actively eliminated from consideration by the decision process. The data, however, does not allow one to say that such an attention shift does not occur before the response is made.

As discussed above, the ability of models to represent the inverse base-rate effect is highly dependent on the rare stimulus being initially presented *after* instances of the common disease. The fact that these are precisely the conditions which give rise to a pre-

response process of 'strategic guessing' means that it is very difficult to say whether the conditions which are responsible for the subsequent manifestation of the inverse base-rate effect emerge purely as a result of post-response rapid attention shifts. More specific experiments are clearly required in order to clarify the nature of both processes.

7.2.4 Summary

Despite the success of the RATM model in being able to represent base-rate effects, important questions about the RATM approach to the study of category learning await attention. The question raised at the start of this discussion, regarding whether the RATM model used in this chapter and that used in the last may be described as the same model, is among the most important of these questions.

The somewhat speculative adjustments made to the learning algorithms for the model in this chapter, in order to 'scale it up' to a four-category structure, represent one difference. This version would, however, scale back down to a two-category model in a way which would, in fact, leave it identical to the RATM model presented in chapter 6.

The most important difference between the two models, as suggested at the start of this discussion, is the removal, from the model presented in this chapter, of the influence of decision uncertainty, or error, from the transition matrix of the sampling system.

The role of this feedback in the representation of the Shepard *et al.* (1961) task difficulties was to enhance the rate of activation of configural representations. This appeared to be required by the model in order to represent early learning rates. As discussed in section 7.1.1.1.1, however, enhancing configural activation would prevent the model from being able to represent the base-rate data described in this chapter. Despite the mixed success of the enhancement strategy in relation to the Shepard *et al.* (1961) tasks, removing the feedback from the model used in chapter 6 is likely to seriously affect early performance of the model on category structures which require configural representations.

The nature of the process required for representation of the base-rate effects is also likely to seriously impair learning of rule-plus exception tasks such as the type V structure. In this case, the attention process is likely to result in a drastic shift of attention away from the 'rule' dimension in the presence of an exception. This will not allow the activation of the three-dimensional representations required to distinguish the exception members of the

categories from the rule-conforming members and will thus severely attenuate learning of these tasks.

As such, it would appear that the shift of attention to error-reducing input components, required to establish the inverse base-rate effect, may not, for some reason, be operating in the same way for learning of the Shepard *et al.* (1961) tasks. While the model uses the same form of representation and equivalent learning algorithms to represent the data from the two paradigms, it would seem that some differences between the nature of the models required to represent these tasks still need to be addressed.

Chapter 8: Overall conclusions

Many specific issues regarding the relative merits and shortcomings of the approaches to modelling described in this thesis have been discussed in the previous chapters. A number of these issues have suggested avenues for further research, which are described in the conclusions of the various chapters. This concluding chapter is, therefore, concerned with evaluating the extent to which the goals of the thesis, outlined in chapter 1, have been addressed. It concludes with a section concerning how the research carried out in this thesis may indicate alternative approaches to other modelling issues in category learning research.

8.1 Goals of the thesis

This thesis has attempted to address a problem, outlined in chapter 1, with current connectionist models of category learning. The problem was described in terms of an apparent 'division' that has emerged in category learning research. This division may be characterised as being between the kinds of experiments which can be modelled using exemplar-based stimulus representations, and the kinds of experiments that cannot.

The nature of this division does not appear to have any basis in a coherent psychological theory. Experiments which can be modelled using exemplar representations are those in which all of the stimuli being learnt about have the same number of dimensions or components. Where the stimulus set includes stimuli with different numbers of components, other forms of representation, such as the configural-cue representation, are required.

The difficulty appears to work both ways in that data from certain experiments to which exemplar-based models have been applied cannot be simulated by models which use a different form of representation. The difficulty here would appear to be of a different nature to that faced by exemplar models with stimuli with different dimensionality.

The problem with the exemplar models is that they lack any principled way of offering any explanations for certain experiments involving stimuli varying in terms of their dimensionality. They lack the capacity to perform certain tasks, which are straightforward for models such as the configural-cue model.

The problem with the configural-cue form of representation is that, while it can carry out many of the tasks to which exemplar models have been applied, its predictions regarding certain aspects of the experimental data are erroneous. This would suggest that it would be easier to alter models using the configural-cue form of representation to allow them to make similar predictions to the exemplar network models, than it would be to alter the exemplar models to allow them to perform tasks which are basic for the configural-cue model.

By taking as its 'starting point' the modelling of the subjective difficulty of the Shepard *et al.* (1961) category learning tasks, the research reported here attempted to address one of the major shortcomings of models making use of the configural-cue form of representation. As discussed by Shepard *et al.* (1961), and described in chapter 2, the task difficulties seem to be related to the number of dimensional values required, before sufficient information to predict the category membership of all stimuli is available. This index of difficulty is not really represented by the basic configural-cue model, which will learn tasks at a rate that is a function of the quantity, frequency and validity of its representations with respect to a task.

While the configural-cue network has all of the information it requires to learn each task, it does not appear to use it in a way that reflects human performance. In order for the exemplar or stimulus generalisation approach to simulate the performance advantage that would accrue for category structures with irrelevant dimensions, Shepard *et al.* (1961) identified a requirement for some form of selective attention process (see section 2.3.2). This requirement would appear to apply equally to the configural-cue representation.

In chapters 5 and 6, several models based on the configural-cue form of representation, but incorporating 'selective' attention processes, were developed. These models all appear to provide qualitative fits to the learning curves reported by Nosofsky *et al.* (1994) that are superior to the variants of the configural-cue model tested by these authors.

The relative merits and shortcomings of these different models are discussed in detail in chapters 5 and 6. The most promising of the models, however, appear to be the dimensional attention models described in chapter 6. While the modular approaches in chapter 5 were able to represent the data, providing them with the ability to do so would

appear to have resulted in the loss of some important aspects of the functionality of the basic configural-cue network (section 5.3).

The dimensional attention models involved the use of a novel conceptualisation of the configural-cue representation as being dependent on the operation of a sequential dimensional sampling process. While this conceptualisation remains somewhat speculative, it does appear to allow selective attention to be implemented using the representation, without obviously losing any of the functionality of the basic model with respect to simple associative learning tasks. This generalisability is particularly true of the ATM and RATM approaches.

These models would appear, for, example, to be able to offer simulations of blocking, transfer from compounds to components, transfer from components to compounds and be able to learn compound-component discriminations in more or less the same way as the basic configural-cue network. Further research with the models would appear to be warranted to investigate in more detail their applicability to these associative learning tasks.

The ATM and RATM models might also be able to offer similar predictions to exemplar models such as ALCOVE (Kruschke, 1992) and its variants regarding the data that these models have successfully simulated. Although the exemplar models obviously have a superior ability, to the models presented in this thesis, to simulate learning about stimuli varying in terms of continuous dimensions, it may be possible to apply adaptations to the form of representation used.

One possibility would be the 'consequential region' approach proposed by Shanks and Gluck (1994) (section 3.2.2.4). More research into the applicability of this approach to the sequential sampling model is required to establish whether incorporating the dimensional attention processes developed here can extend the functionality of the consequential region model.

As discussed in section 3.2.2.4, the configural-cue representation may offer a conceptually preferable account of generalisation between readily discriminable feature-based stimuli such as those used by Shepard *et al.* (1961) and Nosofsky *et al.* (1994), described in section 2.2. The sequential sampling model also offers an account of why

learning gets harder as the number of dimensions required increases. As suggested in section 4.2.2.1, the frequency or arity of the representations alone does not predict this rate.

Exemplar models such as ALCOVE and the GCM appeal to the idea of limited capacity, in relation to specificity or discriminability, to simulate the full effects of the number of relevant dimensions on learning rate (Nosofsky, 1984, Nosofsky *et al.* 1994, p. 366). Exemplar models imply that the rate at which objects themselves get confused for one another increases as a function of the number of dimensions one has to simultaneously pay attention to. The sequential sampling model's account of this increment in difficulty is expressed in terms of the fact that the conditions required for the activation of high dimensionality representations are less probable than those for lower dimensionality representations. Whether this model is generalisable to tasks involving higher dimensionality discriminations remains an area for future research.

The RATM approach, in section 6.3 and chapter 7, appears to show particular promise in relation to its generalisability to other tasks. Incorporating, as it does, a rapid attention shift process, it would appear to possess similar functionality to other rapid attention shift models such as EXIT (Kruschke, in press a and b) and RASHNL (Kruschke & Johansen, 1999). In addition, it is perhaps unique amongst connectionist models of category learning in being potentially able to simulate experiments involving variable dimensionality stimuli, where selective attention and configural representations appear to be necessary. The fact that the RATM model is capable of offering predictions about experiments which involve all of these issues simultaneously, provides an obvious point of contrast between it and other connectionist models which may be worth further examination.

8.2 Further implications: attention 'strategies'

As was discussed in section 3.2.4, it would appear that the way in which generalisation operates is strongly dependent on the tasks in which it is used. As discussed in this section, Tversky (1977) proposed that generalisation may be a flexible process, implemented in ways that take into account different, context dependent, relationships between stimuli and their components.

It was proposed (section 3.2.4) that selective attention processes may be conceptualised as processes that 'control' the way in which generalisation takes place.

Many models, including those presented in this thesis, represent this control in terms of a variation in the influence of particular aspects of stimuli on learning and decision processes.

While the models developed in this thesis appear to offer a way of addressing the representational issues outlined in chapter 1, their development and application to different tasks has highlighted a potential inadequacy with respect to their representation of the apparent flexibility of attentional processes. This was exemplified by the differences, in the role of error in the determination of the effective transition matrix, between the RATM model presented in section 6.3 and the variant used in chapter 7.

For simulating the Shepard *et al.* (1961) and Nosofsky *et al.* (1994) data, there appeared to be some requirement for the model to 'respond' to uncertainty or error by increasing configural activation and associability. For establishing the conditions under which a model will display the inverse base-rate effect, a different response to error or uncertainty seems to be needed. In this case the response to error is to reduce the activation and associability of representations which contribute to error, and increase the activation and associability of representations which reduce error.

One important point, which has emerged, is that it is not necessary to suggest that different forms of stimulus representation mediate the learning of these tasks. The success of models such as ADIT and EXIT in modelling the base-rate effects described in chapter 7, is dependent on them making use of a particular method of implementing selective attention, using a component cue form of representation. While this may suggest that these tasks are learnt in terms of the component cues involved in the stimuli, it has been demonstrated here that there is no need to suggest that configural representations are absent from the system which is learning the task. Indeed, there is no plausible psychological theory which may account for why representation of the stimuli in these particular tasks is restricted to just the individual stimulus components.

The fact that these component cue models succeed, indicates that whatever system does underlie the manifestation of the inverse base-rate effect, it does not make use of configural representations in a way which can be described by the basic configural-cue model. One might suggest that the real challenge posed by this experiment is to account for the manifestation of the inverse base-rate effect, given that people *are* capable of

configurally representing the stimuli involved. By eliminating these forms of representation from models such as EXIT and ADIT, the task of modelling things like the inverse base-rate effect is, arguably, made considerably easier. EXIT and ADIT have no choice but to use only component cue representations of the stimuli.

Other than the differences between representations used in previously successful models, there is no principled reason to suggest that completely different forms of representation underlie performance on the Shepard *et al.* (1961) tasks and the base-rate tasks. There does, however, appear to be a difference in terms of the way in which stimulus representations are *used* in the two tasks. One may describe these different uses of the information afforded by stimuli as representing different processing 'strategies'. While certain theorists have suggested a distinction between 'elemental' and 'configural' processing strategies (Williams, Sagness, & McPhee, 1994, Shanks *et al.*, 1998), the justification for this distinction is generally in terms of the capabilities of extant models using elemental (component cue) and configural representations to simulate the data reported in different experiments.

This simple dichotomy supports assertions that some tasks are best modelled in terms of component cue representations, whereas others may be best simulated using configural or exemplar representations. The apparent necessity, however, to include selective attention processes in models of associative and category learning, makes the distinction between these two types of strategy, as being indicative of different forms of representation, hard to justify.

Selective attention would appear to allow exemplar representations to 'change' into representations of just one or more of the components of the exemplar (sections 3.3.3.1 and 3.3.3.2.2). Configural-cue representations, such as those used in the ASP, ATM, and RATM models, similarly allow flexibility in terms of the dimensionality of the representations which are actually being associated with outcomes. It may be just as appropriate to suggest that different strategies are supported by a process of attention, which operates on the same form of representation to 'allow' it to behave in different ways. As such, differences in apparent strategies being used in different experiments may be describable in terms of the different ways in which attention operates on stimulus representations, within the context of the experiment.

This still leaves the considerable problem of developing models that are capable of simulating the wide variety of strategies that are, apparently, used by people in the performance of category learning tasks.

The RATM model may be described as capturing the differences in the two strategies that appear to be involved in the two types of task modelled in this thesis, in terms of the contribution of error and uncertainty to the determination of the ongoing sampling transition matrix. One might control this contribution using a single parameter to multiply the contribution, prior to its inclusion in the transition matrix. To 'fit' the data for the base-rate experiments, one could set this parameter to zero and set it to a higher value to enable a fit to the Shepard *et al.* (1961) and Nosofsky *et al.* (1994) data.

Without some theory as to why this parameter should vary between the two tasks, little is gained from this approach, other than the knowledge that the stimulus representations are used differently by the model for the different tasks. In addition, the flexibility with which stimulus information is used in learning is unlikely to be representable in terms of this single degree of freedom.

Whether it is appropriate to address the differences between other connectionist models in terms of their being representative of some difference in attentional strategy is, at this point, uncertain. If, as suggested above, selective attention may be describable in terms of a flexible process of control over the way in which representations are used in learning, it might be possible for an appropriate selective attention scheme to 'emulate' other models and, consequently, their performance characteristics.

For this approach to be applied further, considerably more sophisticated models of attention processes would appear to be required. It would appear from the research presented here, and the review of models and research presented in chapter 3, that different strategies may operate for different tasks. In addition, for some tasks, it would appear that people might use information about stimuli in different ways within the same task. A good example of this is provided by the learning of low frequency exceptions to single dimensional rules as described in section 3.3.4.2. The data from this type of experiment is particularly difficult to simulate using dimensional attention models. As was described in section 6.4.1.2, this difficulty is inherited by the dimensional attention models developed in this thesis.

It is also almost certainly the case that there are differences between the strategies employed on the same tasks by different people (Shanks *et al.* 1998, Nosofsky, Palmeri, & McKinley, 1994). Models attempting to simulate, in a general way, associative and category learning processes ought, therefore, to incorporate sufficient flexibility to account for this source of performance variation.

Because there has been a tendency in category learning research to simulate different data using different models, often incorporating different models of stimulus representation, there has been little in the way of research directed at identifying why different models appear to be required in the first place. Despite the, often considerable, differences between the models used to simulate different category and associative learning experiments, the differences between the stimuli used in these experiments are not generally sufficient to justify an assertion that stimuli are represented differently as a function of the experiment.

It would appear, therefore, to be a sensible avenue for future research to explore what it is about the different tasks, which suggests that stimulus information is used differently. What is it about different tasks, which *suggests* that different representations are being used? This would seem to require experiments capable of identifying the kind of variation in task characteristics that informs the way in which selective attention operates in a particular task, or in response to a particular stimulus.

Bibliography

Atkinson, R. C., & Estes, W. K. (1963). Stimulus sampling theory. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology* (Vol. 2, pp. 121-268). New York: Wiley.

Bateson, G. (1979). *Mind and nature: A necessary unity*. Toronto: Bantam Books.

Bartos, P. D., & Le Voi, M. E. (2001). A three-layer configural cue model of category learning rates. In R. M. French & J. P. Sougné (Eds.), *Connectionist models of learning, development and evolution* (pp. 143-152). London: Springer.

Bishop, C. M. (1995). *Neural networks for pattern recognition*. Oxford: Clarendon Press

Bush, R. R. (1960). A survey of mathematical learning theory. In R. D. Luce (Ed.), *Developments in mathematical psychology: Information, learning, and tracking* (pp.125-165). Glencoe, IL: The Free Press.

Bush, R. R. (1965). Identification learning. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology* (Vol. 3, pp. 161-203). New York: Wiley.

Bush, R. R., & Mosteller, F. (1955). *Stochastic models for learning*. New York: Wiley.

Bush, R. R., Luce, R. D., & Rose, R. M. (1964). Learning models for psychophysics. In R. C. Atkinson (Ed.), *Studies in mathematical psychology* (pp. 201-217). Stanford, CA: Stanford.

Carpenter, G. A., & Grossberg, S. (1993). Normal and amnesic learning, recognition, and memory by a neural model of cortico-hippocampal interactions. *Trends in Neurosciences, 16*, 131-137.

Erickson, M. A., & Kruschke, J. K. (1998). Rules and exemplars in category learning. *Journal of Experimental Psychology: General, 127*, 107-140.

Estes, W. K. (1959). The statistical approach to learning theory. In S. Koch (Ed.), *Psychology: A study of a science* (Vol. 1, pp. 380-491). New York: McGraw-Hill.

Estes, W. K. (1991). *Statistical models in behavioral research*. Hillsdale, NJ: LEA.

Estes, W. K. (1994). *Classification and cognition*. Oxford: Oxford University Press.

Estes, W. K., Campbell, J. A., Hatsopoulos, N., & Hurwitz, J. B. (1989). Base-rate effects in category learning: A comparison of parallel network and memory storage-retrieval

models. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15,* 556-571.

Feldman, J. (2000). Minimization of Boolean complexity in human concept learning. *Nature, 407,* 630-633.

Fried, L. S., & Holyoak, K. J. (1984). Induction of category distributions: A framework for classification learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10,* 234-257.

Gati, I., & Tversky, A. (1982). Representations of qualitative and quantitative dimensions. *Journal of Experimental Psychology:Human Perception and Performance, 8,* 325-340.

Gluck, M. A. (1991). Stimulus generalization and representation in adaptive network models of category learning. *Psychological Science, 2,* 50-55.

Gluck, M. A., & Bower, G. H. (1988a). From conditioning to category learning: An adaptive network model. *Journal of Experimental Psychology, General, 117,* 227-247.

Gluck, M. A., & Bower, G. H. (1988b). Evaluating an adaptive network model of human learning. *Journal of Memory and Language, 27,* 166-195.

Gluck, M. A., Glauthier, P. T., & Sutton, R. S. (1992). Adaptation of cue-specific learning rates in network models of human category learning. In *Proceedings of the Fourteenth Annual Meeting of the Cognitive Science Society* (pp.540-545). Hillsdale, NJ: Erlbaum.

Grossberg, S. (1987). Competitive learning: From interactive activation to adaptive resonance. *Cognitive Science, 11,* 23-63.

Hull, C. L. (1943). *Principles of behavior.* New York: Appleton-Century-Crofts.

Jacobs, R. A. (1997). Nature, nurture, and the development of functional specializations: A computational approach. *Psychonomic Bulletin and Review, 4,* 299-309.

Jacobs, R. A., Jordan, M. I., & Barto, A. G. (1991). Task decomposition through competition in a modular connectionist architecture. *Cognitive Science, 15,* 219-250.

Jacobs, R. A., Jordan, M. I., Nowlan, S. J., & Hinton, G. E. (1991). Adaptive mixtures of local experts. *Neural Computation, 3,* 79-87.

Juslin, P., Wennerholm, P., & Winman, A. (in press). *High-level reasoning and base-rate use: Do we need cue competition to explain the inverse base-rate effect?* Manuscript submitted for publication.

Kamin, L. J. (1969). Predictability, surprise, attention, and conditioning. In B. A. Campbell & R. M. Church (Eds.), *Punishment and aversive behavior* (pp. 279-296). New York: Appleton-Century-Crofts.

Karbowiak, A. E. (1969). *Theory of Communication*. Edinburgh: Oliver and Boyd.

Kendler, K. H., & Kendler, T. S. (1968). Mediation and conceptual behaviour. In K. W. Spence, & J. T. Spence (Eds.), *The psychology of learning and motivation* (Vol. 2, pp. 198-244). New York: Academic Press.

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review, 99,* 22-44.

Kruschke, J. K. (1993). Human category learning: Implications for backpropagation models. *Connection Science, 5*, 3-36.

Kruschke, J. K. (1996a). Base rates in category learning. *Journal of Experimental Psychology: Learning, Memory and Cognition, 22*, 3-26.

Kruschke, J. K. (1996b). Dimensional relevance shifts in category learning. *Connection Science, 8,* 201-223.

Kruschke, J. K. (in press a). *Toward a unified model of attention in associative learning.* Manuscript submitted for publication. Available: http://www.indiana.edu/~kruschke/tumaal.html

Kruschke, J. K. (in press b). *The inverse base rate effect is not explained by eliminative inference.* Manuscript submitted for publication. Available: http://www.indiana.edu/~kruschke/elmo.html

Kruschke, J. K., & Bradley, A. L. (1995). *Extensions to the delta rule for human associative learning* [Indiana University Cognitive Science Research Report No. 141]. Available: http://www.indiana.edu/~kruschke/kb95abstract.html

Kruschke, J. K., & Erickson, M. A. (1995). Six principles for models of category learning. Paper presented at the 36th Annual Meeting of the Psychonomic Society, Los Angeles, CA. Available: http://www.indiana.edu/~kruschke/psychonomics95-abstract.html

Kruschke, J. K., & Johansen, M. K. (1999). A model of probabilistic category learning. *Journal of Experimental Psychology: Learning, Memory and Cognition, 25*, 1083-1119.

Kruschke, J. K., & Blair, N. J. (2000). Blocking and backward blocking involve learned inattention. *Psychonomic Bulletin & Review, 7*, 636-645.

Landenowski, S. (1995). Base-rate neglect in ALCOVE: A critical reevaluation. *Psychological Review, 102,* 185-191.

Luce, R. D. (1959). *Individual choice behavior*. New York: Wiley.

Luce, R. D. (1960). The theory of selective information and some of its behavioral applications. In R. D. Luce (Ed.), *Developments in mathematical psychology: Information, learning, and tracking* (pp.5-119). Glencoe, IL: The Free Press.

Luce, R. D. (1963). Detection and Recognition. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.) *Handbook of mathematical psychology* (Vol. 1, pp. 103-189). New York: Wiley.

Mackintosh, N. J. (1975). A theory of attention: Variations in the associability of stimuli with reinforcement. *Psychological Review, 82,* 276-298.

Markman, A. B. (1989). LMS rules and the inverse base-rate effect: Comment on Gluck and Bower (1988). *Journal of Experimental Psychology:General, 118,* 417-421.

McGill, W. J. (1954). Multivariate information transmission. *Psychometrika, 19*, 97-116.

McGill, W., & Quastler, H. (1955). Standardized nomenclature: An attempt. In H. Quastler (Ed.), *Information theory in psychology* (pp. 83-92). Glencoe, IL: The Free Press.

McGill, W. (1955). Isomorphism in statistical analysis. In H. Quastler (Ed.), *Information theory in psychology* (pp. 56-62). Glencoe, IL: The Free Press.

Medin, D. L., & Edelson, S. M. (1988). Problem structure and the use of base-rate information from experience. *Journal of Experimental Psychology:General, 117,* 68-85.

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review, 85*, 207-238.

Miller, R. R., Barnet, R. C., & Grahame, N. J. (1995). Assessment of the Rescorla-Wagner model. *Psychological Bulletin, 117*, 363-386.

Minsky, M., & Papert, S. (1969). *Perceptrons*. Cambridge, MA: MIT Press.

Murre, J. M., Phaf, R. H., & Wolters, G. (1992). CALM: Categorizing and learning module. *Neural Networks, 5,* 55-82.

Nosofsky, R. M. (1984). Choice, similarity, and the context theory of classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10,* 104-114.

Nosofsky, R. M. (1986). Attention, similarity and the identification-categorization relationship. *Journal of Experimental Psychology:General, 115,* 39-57.

Nosofsky, R. M., Kruschke, J. K., & McKinley, S. C. (1992). Combining exemplar-based category representations and connectionist learning rules. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18,* 211-233.

Nosofsky, R. M., Gluck, M. A., Palmeri, T. J., McKinley, S. C., & Glauthier, P. (1994). Comparing models of rule-based classification learning: A replication and extension of Shepard, Hovland, and Jenkins (1961). *Memory and Cognition, 22,* 352-369.

Nosofsky, R. M., & Palmeri, T. J. (1996). Learning to classify integral-dimension stimuli. *Psychonomic Bulletin & Review, 3,* 222-226.

Nosofsky, R. M., Palmeri, T. J., & McKinley, S. C. (1994). Rule-plus-exception model of classification learning. *Psychological Review, 101,* 53-79.

Palmeri, T. J., & Nosofsky, R. M. (1995). Recognition memory for exceptions to the category rule. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21,* 548-568.

Pearce, J. M. (1987). A model for stimulus generalization in Pavlovian conditioning. *Psychological Review, 94,* 61-73.

Pearce, J. M. (1994a). Discrimination and categorization. In N. J. Mackintosh (Ed.), *Animal learning and cognition* (pp. 109-134). San Diego, CA: Academic Press.

Pearce, J. M. (1994b). Similarity and discrimination: A selective review and a connectionist model. *Psychological Review, 101,* 587-607.

Pearce, J. M., & Hall, G. (1980). A model for Pavlovian conditioning: Variations in the effectiveness of conditioned but not of unconditioned stimuli. *Psychological Review, 87,* 532-552.

Pearce, J. M., & Redhead, E. S. (1993). The influence of an irrelevant stumulus on two discriminations. *Journal of Experimental Psychology: Animal Behavior Processes, 19,* 180-190.

Raisbeck, G. (1963). *Information theory: An introduction for scientists and engineers.* Cambridge, MA: MIT Press.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning II: Current research and theory* (pp. 64-99). New York: Appleton-Century-Crofts.

Restle, F. (1955). A theory of discrimination learning. *Psychological Review, 62*, 11-19.

Rosch, E., & Mervis, C. B. (1975). Family resemblance studies in the internal structure of categories. *Cognitive Psychology, 7*, 573-605.

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review, 65*, 386-408.

Rudy, J. R., & Wagner, A. R. (1975). Stimulus selection. In W. K. Estes (Ed.), *Handbook of learning and cognitive processes* (Vol. 2, pp. 269-303). Hillsdale, NJ: Erlbaum.

Rueckl, J. G., Cave, K. R., & Kosslyn, S. M. (1989). Why are "what" and "where" processed by separate visual systems? A computational investigation. *Journal of Cognitive Neuroscience. 1,* 171-186.

Rummelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning internal representations by error propagation. In J. L. McClelland & D. E. Rummelhart (Eds.), *Parallel distributed processing* (Vol. 1, pp.318-362). Cambridge, MA: MIT Press.

Sattath, S., & Tversky, A. (1987). On the relation between common and distinctive feature models. *Psychological Review, 94*, 16-22.

Shanks, D. R., & Gluck, M. A. (1994). Tests of an adaptive network model for the identification and categorization of continuous-dimension stimuli. *Connection Science, 6,* 59-89.

Shanks, D. R., Charles, D., Darby, R. J., & Azmi, A. (1998). Configural processes in human associative learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24,* 1353-1378.

Shanks, D. R., Darby, R. J., & Charles, D. (1998). Resistance to interference in human associative learning: Evidence of configural processing. *Journal of Experimental Psychology: Animal Behavior Processes, 24,* 136-150.

Shannon, C. E., & Weaver, W. (1964). *The mathematical theory of communication.* Urbana, IL: The University of Illinois Press.

Shepard, R. N. (1957). Stimulus and response generalization: a stochastic model for relating generalization to distance in psychological space. *Psychometrika, 22*, 325-345.

Shepard, R. N. (1987). Towards a universal law of generalization for psychological science. *Science, 237,* 1317-1323.

Shepard, R. N. (1994). Perceptual-cognitive universals as reflections of the world. *Psychonomic Bulletin & Review, 1*, 2-28.

Shepard, R. N., Hovland, C. I., & Jenkins, H. M. (1961). Learning and memorization of classifications. *Psychological Monographs, 75* (13, Whole No. 517).

Siegal, S., & Allan, A. R. (1996). The widespread influence of the Rescorla-Wagner model. *Psychonomic Bulletin & Review, 3*, 314-321.

Sternberg, S. (1963). Stochastic learning theory. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology* (Vol. 2, pp. 1-120). New York: Wiley.

Sutton, R. S. (1992). Adapting bias by gradient descent: An incremental version of delta-bar-delta. In *Proceedings of the Tenth National Conference on Artificial Intelligence* (pp. 171-176). Cambridge, MA: MIT/AAAI Press.

Sutton, R. S., & Barto, A. G. (1981). Toward a modern theory of adaptive networks: Expectation and prediction. *Psychological Review, 88*, 135-170.

Tenenbaum, J. B. (1996). Learning the structure of similarity. In D. S. Touretzky, M. C. Mozer, & M. E. Hasselmo (Eds.), *Advances in neural information processing systems* (Vol.8, pp. 3- 9). Cambridge, MA: MIT Press.

Tenenbaum, J. B., & Griffiths, T. L. (in press). Generalization, similarity, and Bayesian inference. Manuscript accepted for publication in *Behavioral and Brain Sciences, 24 (3)*. [Unedited draft available via the World Wide Web at http://www.cogsci.soton.ac.uk/bbs/Archive/bbs.tenenbaum.html].

Tversky, A. (1977). Features of similarity. *Psychological Review, 84,* 327-352.

Wagner, A. R., & Rescorla, R. A. (1972). Inhibition in Pavlovian conditioning: Application of a theory. In R. A. Boakes & M. S. Halliday (Eds.), *Inhibition and learning* (pp.301-336). London: Academic Press.

Widrow, G., & Hoff, M. E. (1960). Adaptive switching circuits. *Institute of Radio Engineers, Western Electronic Show and Convention, Convention Record, 4*, 96-104.

Williams, D. A., Sagness, K. E., & McPhee, J. E. (1994). Configural and elemental strategies in predictive learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20,* 694-709.

Zeaman, D., & House, B. J. (1974). Interpretations of developmental trends in discriminative transfer. In A. D. Pick (Ed.), *Minnesota symposia on child development* (Vol. 8, pp. 144-186). Minneapolis: University of Minnesota Press.